THE HEURISTICS OF STATISTICAL ARGUMENTATION: SCAFFOLDING AT THE POSTSECONDARY LEVEL

by

Teneal Messer Pardue

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Curriculum and Instruction: Mathematics Education

Charlotte

2017

Approved by:

____
Dr. Adalira Sáenz-Ludlow

____
Dr. Pilar Blitvich

____
Dr. Victor Cifarelli

____
Dr. Claudia Flowers

ABSTRACT

TENEAL MESSER PARDUE. The heuristics of statistical argumentation: Scaffolding at the postsecondary level (Under the direction of ADALIRA SÁENZ-LUDLOW)

Language plays a key role in statistics and, by extension, in statistics education. Enculturating students into the practice of statistics requires preparing them to communicate results of data analysis. Statistical argumentation is one way of providing structure to facilitate discourse in the statistics classroom. In this study, a teaching experiment was conducted in which postsecondary students took an introductory statistics course that included formal and informal instruction in statistical argumentation, and a series of tasks designed to support the scaffolding of statistical argumentation. Their work was then analyzed qualitatively to determine how their statistical arguments changed over the course of a semester.

Statistical argumentation is not clearly defined in existing literature. Toward filling this gap in the literature, a formal definition and heuristics of statistical argumentation are developed. Statistical argumentation is defined in this dissertation as a process of justifying a claim using evidence based on data, statistical concepts, and reasoning. Abelson's (1995) five criteria for effective statistical arguments—magnitude, articulation, generality, interestingness, and credibility—are modified to make them appropriate for students at the introductory level. The new criteria consist of three factors: linking to context, articulating results, and making inferences. Students' progress was monitored according to these three criteria.

A constructivist classroom teaching experiment was designed to take into account the institutional curriculum for a first course in statistics at the postsecondary level and special teaching methodology to scaffold statistical argumentation. The teaching experiment was piloted three times to refine the tasks and pedagogy before data was collected for this study in the fourth implementation. A key part of the pedagogy was the development of a classroom culture that supported statistical argumentation. In addition, formal instruction in statistical argumentation took place in four teaching episodes. Each teaching episode followed the same pattern: 1) the teacher-researcher presented a sample argument to the class, 2) students completed arguments in small groups, 3) students completed arguments individually, and 4) students answered reflection questions about their arguments, either in writing or by interview. Students who completed reflections via interview could choose to have the conversation recorded as part of the data collection process. All tasks were the same format: students were asked to answer a research question based on information provided, which consisted of sampling and data collection procedures and results of data analysis in the form of output from a computer software package. The four teaching episodes were based on increasingly advanced statistical content that aligned with the material being covered in class.

The individual statistical arguments and interviews of three students were chosen for full analysis as case studies. The three students were chosen to represent a wide variety of characteristics of statistical arguments. Results show the students made little change in linking to context. However, they showed improvement in the individual aspects of articulation of results: center, spread, distribution, and hypothesis testing.

They were also able to incorporate increasingly advanced statistical content while maintaining or improving in their discussion of previously included content. Students' statistical inference improved over time as well, particularly when hypothesis testing was added to the data analysis; this is not surprising since hypothesis testing is inherently inferential. Student feedback solicited at the end of the teaching experiment indicated that students believed the tasks helped support their learning of statistical concepts and prepared them to interact with statistics in the future.

DEDICATION

This work is dedicated to Ms. Jan Morgan (1957-2012). She inspired her students to believe we could change the world if we cultivated our talents. Her influence remains a guiding force in my life.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Dr. Adalira Sáenz-Ludlow for her guidance and patience both as the chair of my committee and as my professor. I would also like to thank my committee members, Dr. Victor Cifarelli, Dr. Claudia Flowers, and Dr. Pilar Blitvich. I have benefitted greatly from your input.

Special thanks to my students, especially those in this study, for allowing me to learn from you.

I am grateful to my colleagues at Queens University of Charlotte. Ms. Jennifer Daniel offered helpful suggestions on teaching students to write arguments and helped with the process and mechanics in my own writing. Drs. Mike Tarabek and Leina Wu have been supportive department chairs throughout my doctoral work. Dr. Lauren Dimaio has been invaluable as my sounding board for the last chapters.

Thank you to Dr. Lisa Russell-Pinson for her guidance in the writing process.

I would like to acknowledge my five-time undergraduate statistics professor Dr. Steven Patch at the University of North Carolina at Asheville, whose pedagogy of emphasis on statistical language and reasoning shaped my own pedagogy, and led to the topic of this dissertation.

I would especially like to recognize the contributions of my husband Ed, whose love and support in countless ways throughout this process have been nothing short of heroic.

I could not have done this without all of you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

H.G. Wells is quoted as saying, “Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write” (Word Press, n.d.). Indeed, one routinely encounters statistics in the news media, in social situations, and on the job. In the twenty-first century, evidence-based arguments drive decisions in arenas as diverse as business, public policy, and health care. Thus it is imperative that students be prepared to discuss the results of statistical analyses when they encounter them as analysts or consumers. Incorporating evidence-based arguments into the curriculum of an introductory statistics course is one way to engage students in the kind of discourse that occurs in these real-world contexts.

1.1 SUMMARY OF KEY TERMS

The terms in this section will be developed more fully in Chapter 2, but it is useful to have a summary of them here since they will be used throughout this dissertation. These terms are argument, argumentation, statistical argument, statistical argumentation, and criteria of statistical argumentation.

An *argument* is a kind of discourse in which a speaker, called an interlocutor, attempts to convince or persuade an audience of the veracity of some claim. Typically, the interlocutor offers evidence in support of the claim, and reasons that connect evidence together with other evidence or claims. The argument may be directed toward just one person, toward a group of people, or toward a universal audience. Additionally, arguments may be written, oral or even internal to one's own mind.

In contrast to argument, the product of the discourse, *argumentation* is the process of constructing the discourse. The process and the product are very closely related, but the distinction is made for the sake of clarity.

A *statistical argument* is an argument in which the evidence and reasons are based on data and statistical concepts. It may surround any phase of the investigative process, but in this dissertation the claim in the argument is the answer to a research question, the evidence is the results of statistical analysis, and the reasons are statistical concepts that drive interpretation of the statistical analysis. A statistical argument is usually created with the goal of persuading a universal audience but it may also be designed to convince only a select group of individuals.

Statistical argumentation is the process of creating a statistical argument.

In this dissertation, three criteria are proposed to guide statistical argumentation: 1) linking to context, 2) articulating results, and 3) making inferences. The three criteria overlap and interact, but it is useful to consider them separately because they reflect different learning objectives.

Linking to context, where context is the scenario in which the data is situated, requires the argument to include a basic description of the phenomenon being studied, a statement of the value or interestingness of the study, and a discussion of how the results make sense of the phenomenon being studied. Linking to context also requires a consideration of the credibility of the results; i.e., whether the results satisfy common sense and what is currently known about the context.

Articulating results entails discussion of the calculations, charts, and graphs generated from the data analysis. Though many of these, such as measures of center and spread, are considered descriptive statistics, articulation also consists of some inferential results such as confidence intervals and hypothesis tests. These elements should be compared and contrasted in the statistical argument to provide a robust and holistic description of the results.

Making inferences includes all aspects of the statistical argument related to generalizing from the sample to the population. It requires an assessment of whether the data collection procedures are likely to yield a sample that is representative of the population. If any bias is found, the impact it would have on generalizing to the population should be noted. Conclusions such as whether causation can be inferred from the data are also part of making inferences.

1.2 STATEMENT OF THE PROBLEM

Before deciding to use argumentation as a pedagogical tool, the teacher must establish clear learning outcomes. Is the desire for students to “learn to argue” or to “argue to learn”; i.e., does the educator intend for the students to learn to create an argument or is argumentation used as a tool to facilitate learning of concepts? In the case of mathematics, the primary goal of argumentation is for students to refine their understanding of concepts. The argument is less important than the understanding the teacher desires students to gain from it. Similarly, statistical concepts require discussion, i.e., argumentation, to master. However, unlike in mathematics, in statistics it is just as important to explain the meaning of calculations, charts, and graphs as it is to generate them. The interpretation of data analysis, i.e., the statistical argument, is a fundamental product of doing statistics. In statistics education, therefore, it is desired that students will learn to create arguments. Consequently, in this dissertation, the primary focus is on students "learning to argue," although "arguing to learn" is also a goal.

Much of the research on argumentation in education has focused on its use to facilitate learning. Schwarz (2009) notes that creating arguments causes students to think about explanations and clarify their thinking, even when their audience is only imagined. Schwarz also observes the benefit of helping students organize their knowledge—the premises, justifications, and conclusions in the argument. Additionally, there must be cohesion among the premises, justifications, and conclusions of the argument. In the case of introductory statistics, one common difficulty students experience is using correct language and linking the results of statistical analyses to the context of the data (Begg, 1997; Cobb & Moore, 1997). Instruction in the creation of effective arguments may facilitate both of these processes, offering a structure to guide the language students use as well as focusing their attention on the context of the data.

Several studies have focused on the process of learning to create arguments. These studies have shown that argumentation skills can be scaffolded successfully in contexts as diverse as middle school social studies (Nussbaum, 2002) and postsecondary science (Walker and Sampson, 2013) and economics (Zembal-Saul, Munford, Crawford, Friedrichsen, and Land, 2002). These studies show that attempts at scaffolding argumentation as proposed in this study could be of significant interest.

There is a need for more studies on scaffolding argumentation skills in general, especially using scaffolding techniques other than graphic organizers that prompt students to separate parts of their arguments. However, there is a scarcity of research into statistical argumentation. As of this writing, the only studies located were on middle school students' oral arguments specific to a Project-Based Learning environment (Hudson, 2010), and on students' reliance on statistical evidence rather than personal opinions at the middle school level (Osana, Leath, & Thompson, 2004) and in a postsecondary enrichment course outside the required curriculum (Derry, Levin, Osana, Jones, & Peterson, 2000). The present study will examine a set of tasks designed to scaffold statistical argumentation in a postsecondary introductory statistics course, thus taking a preliminary step toward bridging the gap in the literature. The tasks in the proposed study are linked to the introductory statistics curriculum, so they are grounded in course content while simultaneously providing instruction and practice in statistical argumentation. Student work from these tasks will be analyzed qualitatively in order to evaluate the attempts to scaffold students' statistical argumentation.

1.3 RESEARCH QUESTIONS

This study has three research questions:

1. How do introductory statistics students change the way they incorporate the context of a problem into their statistical arguments over the course of a semester as they complete tasks designed to scaffold their argumentation skills?
2. How do introductory statistics students change the way they articulate the results of data analysis into their statistical arguments over the course of a semester as they complete tasks designed to scaffold their argumentation skills?
3. How do introductory statistics students change the way they make inferences in their statistical arguments over the course of a semester as they complete tasks designed to scaffold their argumentation skills?

For the first research question, the working hypothesis is that students will increasingly ground their statistical arguments in the context of the data. Ideally students will use the results of the data analysis to make sense of the context in a meaningful way.

For the second research question, the working hypothesis is that students will become more adept at combining the results of various charts, graphs, numerical summary measures, and hypothesis tests to articulate an overall description of the data.

For the third research question, the working hypothesis is that students will better identify bias in sampling procedures and adjust the claims and conclusions made in their statistical arguments accordingly.

1.4 SIGNIFICANCE OF THE STUDY

This study has implications for the practice of teaching statistics. Pedagogically, if the study reveals that focused instruction in scaffolding argumentation can facilitate the creation of better statistical arguments, then it will point to the potential for an effective instructional technique to be developed. Further, it may be possible to show that students at the introductory level are capable of constructing basic arguments, whereas Abelson's (1995) belief was that a more advanced knowledge was necessary. This study is consistent with the recommendations of a joint committee of the American Statistical Association and the Mathematics Association of America that introductory statistics courses should focus on data analysis and concepts. Encouraging argumentation is a natural way to focus students' attention on concepts as well as aid in data analysis. Ideally, this study will begin to reveal a way to scaffold students' statistical argumentation so that they can effectively communicate the results of data analysis and better understand the course concepts learned in the introductory statistics course.

Since statistical argumentation is a fairly new concept, research is needed to determine its benefits, how it might be taught effectively at various levels of statistical expertise, ways of assessing it, and what it might reveal about student understanding of statistical ideas. This study will ideally contribute to this emerging topic of research and point to some further areas for investigation.

CHAPTER 2: THEORETICAL FRAMEWORK

The theoretical framework is divided into five sections: differences between mathematical reasoning and statistical reasoning, argumentation, statistical argumentation, the sociocultural theory of learning, and scaffolding.

In the first section, elements of statistical reasoning that are different from mathematical reasoning are discussed, particularly the fundamental role of context, the importance of language, and the use of subjective decision-making, all of which are unique to statistics.

In the second section, the origins of argumentation are recounted, starting with Aristotle and then continued centuries later by Toulmin (1958) and Perelman and Olbrechts-Tyteca (1958). Definitions of argument and argumentation are developed.

In the third section, definitions of statistical argument and statistical argumentation are developed. Abelson’s (1995) conception of effective statistical arguments is explained. Finally, based on the work of Abelson, a model of effective statistical arguments in the introductory statistics curriculum is presented.

In the fourth section, the sociocultural theory of learning, particularly based on the writings of Vygotsky and later Luria, is described. The Zone of Proximal Development (ZPD) and metacognition, which are fundamental to this study, are reviewed.

In the fifth and final section, the idea of scaffolding instruction is recounted. The particular challenges of scaffolding instruction for a group are discussed.

2.1 DIFFERENCES BETWEEN MATHEMATICAL REASONING AND STATISTICAL REASONING

There is a strong relationship between mathematics and statistics. They are generally taught by the same educators, often in the same academic department. Many educators apply the same teaching philosophy and pedagogy to statistics as they do to mathematics classes. However, researchers in mathematics and statistics education have identified three fundamental ways in which they are different: 1) in statistics, some of the decisions that must be made are based on subjective criteria which are matters of judgment rather than analytical reasoning; 2) in statistics, use of language takes on greater importance because of the need to communicate results; and 3) in statistics, the context of a problem plays a fundamental role in guiding the decisions and interpretations. These differences point to the need for statistics to be taught using another approach and for statistics education to be seen as a separate field of research.

The first and perhaps the most fundamental difference between statistical reasoning and mathematical reasoning is that while mathematics is objective, statistics requires a degree of subjectivity. As Perelman (1982) points out, mathematics is a formal system that is subjected to restrictions in an attempt to eliminate ambiguity. However, in statistics, while calculations may be precise, other elements involve some ambiguity such as whether or not to discard an outlier, which of two models is preferable, or how to interpret the results. Every step in a mathematical problem or proof can be justified based on mathematical rules or principles that have themselves been proven. However, the person conducting a statistical analysis must at times make decisions without clear-cut guidelines.

The second is that data analysis requires the statistician to use his or her own judgment. Here it is imperative for the researcher to justify those decisions that involve subjectivity in judgment. Indeed, Begg (1997) proposes that one way in which statistics is different from mathematics is the role of communication. Whether implicit or explicit, every statistical analysis is prepared with a goal of communicating the results to a "consumer" who will use the conclusions to answer questions about a real-life scenario. Thus it is crucial that students learn to discuss statistics and justify their conclusions so that in their future coursework and careers, they are prepared for the role of statistical analyst or consumer.

The third and final difference between mathematics and statistics is the role of context. Cobb and Moore (1997) suggest that in mathematics, the ultimate goal is to strip away context so that the structure of problems is the focus. By contrast, the context provides meaning to statistics. A concept such as standard deviation or t test is meaningless without a situation in which to apply it. Problems abstracted from context are analyzed in statistical theory courses, but the branch of statistics devoted to the study of statistical structures is called mathematical statistics, which further suggests that the distinction proposed by Cobb and Moore is a valid one.

The type of mindset and skills required to be successful in statistics have become known as statistical literacy, reasoning, and thinking. Ben-Zvi and Garfield (2004) describe each of these in detail. Statistical literacy refers to an individual's understanding of basic vocabulary, symbols, probability, data description, and data display. Statistical thinking involves a broader understanding of variation, sampling, inference, and interpretation. An individual who engages in statistical thinking is able to critique studies and use the context of the data in interpreting results of data analysis. Statistical reasoning refers to a person’s ability to summarize, explain statistical processes, and interpret results, while connecting concepts to one another.

The unique thinking and reasoning skills required for statistics necessitates that pedagogy for statistics courses be different from that of mathematics courses. Cobb and Moore (1997) relate the results of a joint committee of the American Statistical Association (ASA) and the Mathematics Association of America (MAA) designed to discuss the curriculum and approach in introductory statistics courses. The committee concluded that “any introductory course should take as its main goal helping students to learn the basics of statistical thinking” (p. 803). The committee recommended that, unlike mathematics where theory is a main focus, in statistics the emphasis should be on data and concepts rather than on theory and hand calculations.

The research to be conducted for this dissertation is consistent with the observations of Cobb & Moore and Begg, and the recommendations of the ASA and MAA joint committee. Giving students real-life scenarios with which to work preserves the central role of context. Focus on the process of justifying and communicating decisions, either to oneself or others, prepares students to communicate statistical ideas in their future coursework and careers. Finally, the tasks proposed in this study concentrate students’ attention on interpretation of results, while calculations are completed using computer software, thus satisfying the recommendation of the joint committee.

2.2 ARGUMENTATION

Since the general theory of argumentation is the foundation of statistical argumentation, the definition of argument itself must be clarified before proceeding further. Argumentation and rhetoric as a field of study dates back to Aristotle, who described a persuasive argument as “one that persuades the person to whom it is addressed” (as cited in Perelman, 1982) and which, unlike analytical reasoning, “derives its value from its action upon the mind of some person” (as cited in Perelman, 1982). Both analytical reasoning and rhetoric start with a set of theses that are generally accepted and then explain how the existing theses lead to a new thesis. However, Aristotle saw argumentation as different from analytical reasoning because it aims to justify opinion rather than demonstrate proof. Argumentative discourse fell out of favor by the end of the 16th century, but experienced a resurgence in 1958 when two seminal works were published: Stephen Toulmin’s *The uses of argument* and Chaim Perelman and Lucie Olbrechts-Tyteca’s *The new rhetoric: A treatise on argumentation*. These two works have formed the basis of argumentation theory.

The definitions in this section indicate a distinction between argument and argumentation. An argument is usually defined as the discourse itself while argumentation is the process of creating arguments. Rowland (1987) uses the word *argument* to describe both the discourse and creation of the discourse, maintaining that the product and process are inseparable. For the sake of clarity, in this paper the word *argument* will refer to the product and *argumentation* will refer to the process of constructing the discourse, with the understanding that the product and process are very closely related.

Toulmin (1958) proposes a diagram to describe the structure of an argument, which is useful in understanding the concept of argument as well as in dissecting particular examples of arguments. Other models which followed from that of Toulmin are similar and follow the same basic structure as above. Toulmin’s model consists first of a claim that is supported by facts called data or evidence. Justification for how the data supports the claim is called a warrant, and if the warrant itself requires justification, it is called backing. Finally, any limitations or restrictions to the claim are called the rebuttal. The diagram proposed by Toulmin is shown below.

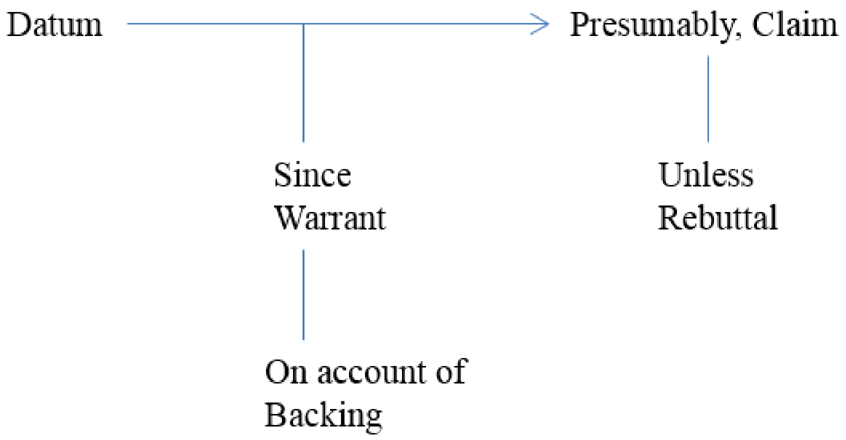


Figure 1. Toulmin Model. Adapted from *The Uses of Argument* (p. 104), by S. E. Toulmin, 1958, London: Cambridge University Press. Copyright 1958 by Cambridge University Press.

While Toulmin’s primary purpose was to provide a structure to individual arguments, Perelman and Olbrechts-Tyteca (1958) and later Perelman (1982) characterize not only the arguments themselves, but some key features of arguments, what they describe as being different from proof in that they are not self-evident and require some irrational elements. The goal is for an interlocutor to “increase the adherence of the members of an audience to theses that are presented for their consent” (Perelman, 1982, p. 9).

Perelman and Olbrechts-Tyteca (1958) put considerable emphasis on the audience, those people who the speaker is trying to influence with the argument. As they point out, the audience is not merely a set of spectators but a key factor in the creation of the argument. The interlocutor needs to consider the audience at every phase of the argument, by choosing premises that the audience already accepts, choosing evidence and reasons that will be compelling to the audience, toward a conclusion that the audience will accept and see as important. The audience may be any number of people, as few as one person or as many as all of humanity. There are two special cases which warrant mention. The first is that the audience may consist of the interlocutor him- or herself engaging in an internal dialogue, like what happens when students reflect on ideas and evaluate whether claims are supported by known facts and reasons. The other is the universal audience, which consists of all rational beings. This is the intended audience for most statistical arguments. Perelman and Olbrechts-Tyteca refer to arguments that appeal to the universal audience as *convincing*, whereas arguments directed to smaller groups as *persuasive*. Convincing arguments require a stronger standard of evidence, closer to analytical reasoning, than persuasive arguments.

Perelman and Olbrechts-Tyteca also characterize several types of arguments. Of particular interest here is what they call a *quasi-logical argument*, which is an argument that often relies on probability. By using probabilities to support a claim, they assert that the elements become more easily comparable to the audience. The challenge of using probabilities to support a claim is that it can become very complicated to take into account all relevant factors in computing the probabilities. The argument will be weak if the interlocutor does not address concerns about the basis for the probabilities that are used in the argument.

Since 1958, argumentation has been studied in a variety of contexts, developing so much that there are currently several journals dedicated entirely to the subject. It is so widely used that scholars do not all agree on one definition. A scan of the literature was conducted in an attempt to find a definition of argumentation that represents a consensus of scholars of argumentation. Forman, Larreamendy-Joerns, Stein, and Brown (1998) defined argumentation broadly in a mathematical context of “the intentional explication of the reasoning of a solution during its development or after it” (p. 531) while Hitchcock (2002) proposes that argument is “a spoken discourse or written text whose author (the arguer) seeks to persuade an intended audience or readership (the Other or the Others) to accept a thesis by producing reasons in support of it” (p. 289). The variety of definitions is so great that Rowland (1987) was prompted to observe “some theorists believe that others are not even studying argumentation” (p. 141). Most of the definitions of argument or argumentation in a review of literature had six common elements: claim, persuasion, evidence, reason, and debate.

In his article “On Defining Argument,” Rowland (1987) develops a definition of argument based on an analysis of the work of other theorists. His definition of argument as a product is “that type of discourse in which reasons and evidence are presented in support of claims” (p. 148). He defines argument as a process (what is being called argumentation here) as occurring when “an individual tests his or her claims against the claims of other arguers" (p. 148). These definitions include all the elements described above except for persuasion, which Perelman and Olbrechts-Tyteca (1958) assert is the "goal of all argumentation" (p. 45). Rowland's definition includes a different idea (though the two are not mutually exclusive), that of testing claims against one another, presumably in an attempt to determine the validity of each claim. Rowland goes on to offer an alternative to debate with another person: "a single individual may engage in what might be called an internal dialectic, in which he or she tests every claim against counter claims" (p. 148). This kind of internal, reflective argumentation is especially important in education, where students learn by interacting socially with the goal of internalizing the concepts and processes. Rowland's definition is the one adopted for this dissertation with one exception. The goal of argumentation may be either, as Rowland postulates, to test the validity of arguments, or it may simply be to persuade another person or group of people.

To summarize, arguments begin with a claim, which may be explicit or implicit. The claim is the idea, also called a proposition or a proposal, which the arguer desires to establish. The goal of the argument is for participants to come to a shared understanding. Evidence is given in support of the claim. Reasons are given which connect evidence together with other evidence or to claims. Often it involves some debate or counter-argument.

2.3 STATISTICAL ARGUMENTATION

The obvious subject matter of statistical argumentation is statistics, but further detail must be supplied. Pfannkuch and Wild (2004) identify five phases of the investigative cycle: 1) defining the problem, 2) planning, 3) data collection, 4) analysis, and 5) conclusions. At each phase of the investigative process, the statistician chooses from among several possible courses of action. If-at any point in the process, the statistician cannot justify his or her decisions, the credibility of the entire study may be called into question. The first phase is defining the problem. The statistician chooses a research question and creates operational definitions of key terms. In the second phase, planning, the researcher designs an experiment, survey, or other measurement tools. Sampling techniques are chosen. Even when the statistician is using data that has been created and collected by someone else, he or she must still choose which dataset to use. The third phase is data collection, which includes not only the actual process of collecting data but also cleaning and managing the data. During the fourth phase, analysis, the statistician uses a variety of charts, graphs, numerical summary measures, and hypothesis tests to glean information from the data. The fifth and final phase is conclusions, which calls for the statistician to interpret and communicate the results of the analysis. Statistical argumentation must therefore include any course of action at one or more of these five stages.

These five stages as defined by Pfannkuch and Wild correspond to key elements in a statistical argument. At each phase, the interlocutor needs to justify decisions made in the analysis. In some instances, particularly when elements of the investigative cycle are performed by another researcher, these elements will need to be evaluated rather than justified; any shortcomings will be incorporated into the conclusion as limitations of the study. In the first stage of a statistical argument, the interlocutor explains or restates the research question. In the second phase, sampling procedures are discussed, with attention paid to any sampling bias. In the third phase, data collection procedures are discussed, and any issues that might cause the data to be compromised are noted. In the fourth phase, characteristics of the sample are reported; usually this takes the form of charts, graphs, and numerical summary measures. Inferential procedures such as hypothesis tests and confidence intervals are also reported as part of the fourth phase. In the fifth and final phase, conclusions are presented. Conclusions include an answer to the research question. Any problems noted with sampling and data collection procedures, or any assumptions not met for inferential procedures need to be synthesized into the conclusion as limitations.

A common characteristic within a field of argument is a shared purpose. In the case of statistics, analyses are conducted for two main purposes. First, the statistician sets out to prove a claim and find information to support it. Since the goal is to support the claim, contradictory information is likely to be ignored. Second, in contrast to the first, a researcher wants to test a hypothesis, i.e., a claim, not for the purpose of supporting it but rather to determine whether it can be substantiated from the data. This difference is described by Abelson (1995), who calls the first analysis *conservative* when it is conducted to support a claim. In contrast, he calls the second analysis *liberal* when it is conducted to test a claim. In general, the goal of the analysis is to impact the decisions and resulting arguments at each phase of the investigation. For example, if the goal is to support a claim, then the decisions are likely to be defended with no willingness to consider counterarguments, while the statistician trying to test a claim may be more apt to encourage discussion and even adjust his or her choices and arguments as a result.

Taking these elements into account, the following working definition is proposed in this study: statistical argumentation, as a part of the field of argumentation, is a process of justifying a claim using evidence based on data, statistical concepts, and reasoning. The goal, in general, is to create an argument that persuades a universal audience but also that convinces a select group of individuals.

2.3.1 EFFECTIVE STATISTICAL ARGUMENTS: ABELSON'S MODEL

Though statistical arguments are common, little has been written about how statistics should be used to construct convincing or persuasive arguments. One notable exception is Abelson's (1995) *Statistics as principled argument,* in which the author proposes five criteria—magnitude, articulation, generality, inference, credibility—that cause an argument to be compelling. These criteria may provide useful guidelines for helping students create statistical arguments. While statistical arguments may surround any phase of any investigative process, Abelson addresses arguments only regarding the conclusions of data analysis. However, to support the conclusions it is necessary to defend decisions of other parts of the investigative process as well.

Abelson describes two possible explanations that may be claims in statistical arguments: the data could be the result of a random process or there could be some systemic explanation causing departures from randomness. These two claims, randomness versus systemic explanation, relate closely to the paradigm of null hypothesis testing, in which randomness is assumed unless sufficient contradictory evidence can be found. However, Abelson's characterization is general enough to allow for other types of evidence to be used, even a basic statement of group means or a graph demonstrating the claim. Indeed, Abelson criticizes null hypothesis testing as relying too heavily on single studies, yielding results that are statistically but not practically significant, and imposing upon the data a false dichotomy. All of these criticisms point to the importance of the data analyst's ability to look beyond the mere results of a hypothesis test and express a more nuanced argument.

Maintaining that "[Data analysis]…should make an interesting claim; it should tell a story that an informed audience will care about, and it should do so by intelligent interpretation of appropriate evidence from empirical measurements or observations," (p. 2) Abelson proposes the Magnitude, Articulation, Generality, Interestingness, Credibility (MAGIC) criteria for constructing an effective statistical argument. Each of the five elements will next be described in detail.

Magnitude, the first component of the MAGIC criteria, requires the data analyst to explain not only if there is a systemic factor at play in the data, but how strong the result is. Usually, it requires stating more than merely the result of a hypothesis test. While the p-value itself is one indication of effect size, with smaller p-values being stronger evidence for the alternative hypothesis, p-values are strongly dependent upon sample size. Consequently, with a large enough sample, a very small correlation or difference in means may result in rejection of the null hypothesis, despite the lack of any meaningful relationship. While the p-value of a test provides information about the statistical significance of the result, effect sizes such as the correlation coefficient provide information about the practical significance.

The second component, Articulation, refers to the detail with which relationships are hypothesized, described, and qualified. Abelson labels the description of detailed research results as *ticks* which can become part of the general knowledge of a field when they gain wide acceptance by scholars of the discipline. Ticks should be accompanied by *buts*, statements that constrain ticks. Buts describe situations in which the postulated relationships have not been demonstrated or need further investigation.

The third factor, Generality, is related to articulation. Generality refers to the conditions in which the results can be expected to hold. Most studies take place in particular contexts, with participants that are chosen from a narrow population. Researchers must describe any limitations in making inference to the population of interest. To minimize limitations to generality, replications of the study under varying conditions can be conducted. Similar studies with variations can also be cited in the results to support generality.

Interestingness, the fourth part of the MAGIC criteria, has a subjective element but could take several forms. The results may have real-world applications that cause the study to be important. It could have implications for shaping theory in the discipline or offering limitations or generalizability to other studies that have been conducted. A result that is surprising because it violates expectations or intuition creates an interesting argument.

Though a surprising result satisfies interestingness, it must not be so counterintuitive as to jeopardize credibility, the final element of a compelling argument. According to Abelson, there are two primary ways a claim might violate credibility: it contradicts common sense or strong beliefs, or it is based on poor methodology. The job of the statistician, then, is to justify the methodology used and give explanations for any results that may seem intuitively incorrect.

Abelson’s description of the possible statistical arguments that might be made—randomness versus departure from randomness—are applicable to statistical methods ranging from the basic ones included in the middle school curriculum up to more sophisticated methods used by professional statisticians. His MAGIC criteria to describe effective statistical arguments are applicable in a wide variety of methods as well, though they work best with more advanced methods. Abelson's work forms the basis of what will be defined as a standard of statistical argumentation. As such, it is a key framework for this dissertation.

2.3.2 AN IMPLEMENTATION OF ABELSON'S MODEL OF STATISTICAL ARGUMENTATION

During the study, students will be instructed in an ideal of statistical argumentation based on Abelson's model. Abelson describes his intended audience for *Statistics as principled argument* as students who are experienced with statistical concepts. He relates, "It has been my observation that most students do not really understand statistical material until they have been through it three times—once for exposure, once for practice, and once for the dawn of genuine insight. This book is designed for the third pass-through…" (p. xiv). Since introductory statistics students usually have little or no prior experience with statistical methods when they begin the course, Abelson's work will need to be adapted to the appropriate level. The following paragraphs describe the modifications that are made in this dissertation.

For use in the introductory statistics course, Abelson's five criteria for arguments are all included, but reorganized so that the criteria fit the course curriculum, align with the investigative process, and address the needs of introductory students. In this new organization, there are three criteria: linking to context, articulating the results, and making inferences. Criteria are presented to the students in increasing detail throughout the semester as relevant concepts are covered. The criteria are described in the order in which they occur through the investigative process; this order is more logical to the students and increases the likelihood that they will consider all important elements.

The first part of the criteria of statistical argument as presented in the introductory statistics course is linking to the context of the data. Context includes Abelson's conception of interestingness as well as credibility of the results, but is even broader than those two criteria. Context is an inextricable part of a statistical argument and is present in every phase of the argument. It requires a general knowledge and sense of the practical implications of the argument. First, linking to context is present in the consideration of whether the research question and data collection procedures make sense for the scenario being discussed. For example, if (as in task 1a, see Appendix A), academic advisors at a college conduct a survey of whether students have cheated academically, it would include pointing out that participants have motivation to lie because the survey is not anonymous and they would likely not be comfortable to disclose to their academic advisor if they had cheated in school. This is a connection to the phenomenon being studied. Second, linking to context is present in articulating the results in the scenario of the problem. For example, the statement, "City 1 has a smaller mean commute distance of 15.2 miles and city 4 has a larger mean of 20.1 miles." (given in the sample argument of Task 2, see Appendix A) situates the results in the context whereas an alternate statement "Group 1 has a mean of 15.2 and group 4 has a mean of 20.1." does not. Third, linking to context is present in the consideration of the credibility of results. For example, if 85% of teens surveyed reply (as in Task 1a, see Appendix A) that they do not favor suffrage, it indicates that the sample is not representative of the population and/or the question is confusing. Thus the context criteria requires students to link the analysis to the context at the planning phase as well as after the calculations have occurred.

The second criteria of this proposed version of Abelson's criteria is articulating the results. These are the *ticks* and *buts* Abelson described. Much of the introductory statistics course focuses on the articulation aspect of statistical arguments. Three main elements are included under articulation of results. First, descriptive statistics, including measures of center and spread, need to be discussed. Second, the distribution of the data needs to be evaluated, as evidenced by histograms, boxplots, and frequency tables. Finally, appropriate hypothesis tests need to be implemented and interpreted. Although hypothesis tests are inferential statistical methods, they are included here as part of articulating results rather than making inferences. The reason is that hypothesis tests are part of the analysis, and they provide additional information about the data.

For the purpose of introductory statistics, magnitude is included as a part of articulation rather than as its own category. Effect size is rarely covered as part of the introductory statistics curriculum. However, effect size can be discussed by introducing Cohen's d, a measure of effect size for one-sample t tests and independent samples t tests. For the one-sample t test, Cohen's *d* is calculated as $d = \frac{|\bar{x}-\mu_0|}{s}$, where $\bar{x}$ is the sample mean, μ_0 is the hypothesized value for the population mean, and s is the standard deviation of the sample. For the independent samples t test, Cohen's *d* is calculated as $d = \frac{|\bar{x}_1-\bar{x}_2|}{s}$, where $\bar{x}_1$ and $\bar{x}_2$ are the means of the two groups and *s* is the standard deviation of one of the samples (Cohen, 1988). Since in the independent samples t test used in the course, the assumption is not made that the group standard deviations are equal, students are taught the conservative approach of using the larger one. Cohen expressed reservations about establishing guidelines of what constitutes “small,” “medium,” and “large” effect size because it depends in part upon the particulars of each study. However, he suggested 0.2 as a small effect, 0.5 as a medium effect, and 0.8 as a large effect. Effect size is a simple concept and a simple calculation, and is used to add detail to the results of hypothesis tests.

The final norm of statistical argumentation for students is making inferences, a broadened conception of the generality criteria. Inferences are conclusions made about a population based on a sample. The first part of making inferences is thus considering how well the sampling procedures yield a representative sample of the population. Any bias in sampling will directly impact the conclusions that can be made. Sampling processes are included in the introductory statistics curriculum as part of a discussion of critical thinking based on Huff and Geis’s (1954) classic book *How to lie with statistics*, which details some common ways that statistical information can be misused and misinterpreted. One focus in the critical thinking part of the course is on the validity of making inferences based on a given sample. The second and final part of making inferences is stating the conclusions of the analysis. This requires that all the information be synthesized into a clear answer of the research question. Any limitations on the conclusions based on problems with the sampling and data collection procedures need to be noted. Some recommendations may be made for modifying the study so the research question may be better answered.

These three elements of statistical argumentation—linking to context, articulating results, and making inferences—include Abelson’s MAGIC criteria but are tailored for use with the introductory statistics curriculum. They are general enough to encompass statistical arguments at the basic level as well as more advanced levels, but they offer direction for both the creation and evaluation of statistical arguments.

2.4 SOCIOCULTURAL THEORY OF LEARNING

This study is grounded in the sociocultural theory of learning, especially as described by Vygotsky. Self-regulation is an essential component of Vygotskian sociocultural theory. Vygotsky claimed that self-regulation has origins in social life. This study examines the effect of a set of social activities on students' ability to make sense of and communicate the results of statistical analysis, a self-regulatory behavior. Because the interventions in this study—argumentation combined with oral and written dialogue—are grounded in social activity to scaffold instruction, they are fundamentally Vygotskian in nature.

The argumentation exercises are also based on Vygotskian theory. The instructor first models argumentation during lessons, thus enculturating students into the language of argument. Next, students will practice creating arguments with help from one another while the instructor guides the students to be precise in their arguments. As students internalize the process of creating claims and justifying them with data, their thinking becomes self-regulated. That is, they hold themselves accountable to the process of argumentation that has previously been imposed upon them. They are able to do alone what they could previously do only collaboratively, which is the essence of Vygotsky's idea of the Zone of Proximal Development (ZPD) (Moll, 1990).

Building on the work of Vygotsky, Luria theorized that in order to learn a concept, students must be able to manipulate the concept voluntarily (Diaz, Neal, &

Amaya-Williams, 1990). In this case, argumentation is a way for students to make conscious use of formal reasoning in the subject matter. Vygotsky specified five forms of mediation: collaboration, direction, demonstration, leading questions, and introducing the initial elements of activities (Moll, 1990). All these modes of mediation will be incorporated throughout the semester, as the instructor employs direct instruction, provides examples, then guides and facilitates written discourse (Moll, 1990).

Voluntary manipulation of a thought process is also called metacognition, an important element of this study. Argumentation is essentially a formalization of reasoning that gives students an opportunity to reflect upon their thought. By introducing this process of focusing on the thought process itself, students are better able to manipulate the process and improve upon their success in creating data-based arguments. Since written expression of mathematical reasoning has been shown to increase metacognitive behaviors (Pugalee, 2004), the hypothesis in this study is that further metacognition and thus more success may be developed when students record their arguments in writing.

Yackel and Cobb (1996) concur, claiming that explaining one’s reasoning clarifies aspects of the individual’s mathematical thinking that may not be apparent to other people. Further, Yackel and Cobb assert that when students are encouraged to focus and reflect on the explanations, they begin to understand argumentation as a process as well as argument as a structure. Reflecting on the arguments gives further basis for metacognition. Students learn to evaluate for themselves what arguments are acceptable.

2.5 SCAFFOLDING

The term “scaffolding,” first coined in 1976 by Wood, Bruner, and Ross, is a metaphor which refers to an instructor’s role in supporting student progress along the ZPD. Scaffolding as they described it is the process by which a tutor helps a novice accomplish a goal by means of assistance that is withdrawn as the novice becomes able to achieve the goal independently. They identified several ways in which scaffolding may occur: the tutor models the task, simplifies the task to make it manageable, maintains the novice’s focus in a productive direction, and marks important features of the task. Throughout the process, the tutor provides feedback and offers enough help to minimize frustration while allowing the novice to work as independently as possible.

In the years since it was first described, scaffolding has become common terminology in educational literature. Researchers have identified additional instructional activities that occur in scaffolding. Hogan and Pressley (1997) point out that scaffolding must start before the tutor begins interacting with the student, with the selection of a task appropriate to the student’s current level of performance and consistent with curriculum goals. Continually assessing the understandings and needs of the learner allows the tutor to give feedback, which may consist of pointing out progress and behaviors that lead to success. When the student is successful, the tutor may explicitly restate the steps followed and concepts learned. Roehler and Cantlon (1997) add that an important step is for the tutor and learner to establish a shared understanding of the task. Pressley, Hogan, Wharton-McDonald, Mistretta, and Ettenberger (1996) point to the importance of errors, which can become an important tool in scaffolding. A tutor may even make mistakes on purpose to allow the learner to suggest alternatives.

In a review of the literature on scaffolding, van de Pol, Volman, and Beishuizen (2010) found that authors universally consider instruction which is tailored to the needs of the individual as a key element of scaffolding. Accordingly, most studies of scaffolding occur with one student at a time. When scaffolding is attempted with a group of students, it becomes a far more complicated endeavor. Hogan and Pressley (1997) described some of the challenges of scaffolding in a classroom setting and suggested ways to address those challenges. In a group of students, each student may have a different level of understanding and competence, and the ones who need the most help are least likely to ask for it. Students may also have a variety of communication styles due to cultural and linguistic diversity. The inability to interact with each student individually means the instructor cannot assess each student’s progress, making it even more important that the instructor have a thorough grasp on the material and a deep level of insight into how students tend to interact with the course content. Van de Pol, *et al*. (2010) recommend that giving students in small groups tasks to complete with a tool such as cue cards allows the students to scaffold one another. The instructor can take turns working with each small group, offering the group the same kinds of scaffolding that might be provided to one individual student. Another recommendation is that in whole-class discussions, the instructor responds to student contributions with the goal of generating further questions and thought process rather than evaluating as correct or incorrect. Even when interacting with the class as a whole, scaffolding behaviors can occur. The instructor should offer support to allow the group to accomplish the task of generating a shared understanding and ability to manipulate a concept, while gradually withdrawing the support as the class is able to function independently.

CHAPTER 3: LITERATURE REVIEW

This literature review is divided into two sections. The first section presents research on scaffolding argumentation. The second section describes studies that relate to statistical argumentation, with emphasis on research relating to the three norms of statistical argumentation: linking to context, articulating results, and making inferences.

3.1 SCAFFOLDING ARGUMENTATION

Several studies have shown that argumentation skills can be scaffolded successfully. One common technique is the use of a form that requires students to create arguments by specifying the claim, backing, warrants, and supporting data. Nussbaum (2002) employed one such form with a group of sixth grade students in two social studies classes. Over a thirteen week period, students formed a classroom government and designed a model city. Students were presented six scenarios during the course of the study in which they were required to make choices of what to fund and what policies to implement in their model city, and they were asked to justify their choices (claims). The form used to scaffold argumentation included blanks for students to fill in their choices, evidence to support their choices, how the evidence related to their choices, and a restatement of their claims. Nussbaum found that students used more evidence over time to support their claims. Not surprisingly, students struggled with the more difficult task of explaining how the evidence supported their choices, but students improved on that task through the course of the study as well. Students with better language skills (as measured by an achievement test) generated better arguments, though students with lower scores successfully used the scaffolds.

Zembal-Saul, Munford, Crawford, Friedrichsen, and Land (2002) used a similar scaffolding technique embedded in a software package, which gave students boxes to fill in for claims and evidence, then required them to rate their arguments and explain their rating. Participants, who were pre-service secondary science teachers, completed a series of arguments using the software in a unit on natural selection. Findings included that the scaffolds did support the development of evidence-based arguments but that weaknesses in argumentation persisted at the end of the unit.

Cho and Jonassen (2002) also used software with a similar scaffolding technique. Participants were students in an undergraduate introductory economics class. They were assigned to work problems in groups of three, communicating via an online discussion board. Using random selection, half of the groups were given the scaffold and the other half were not. Cho and Jonassen found that groups who were given the scaffold focused more on claims and supporting data because they are most emphasized in the software. Based on the scores of two raters who were blind to the treatment given, the authors were able to conclude that the arguments of the scaffolded groups were overall better than those of the control groups.

In a study of argumentation scaffolding in a General Chemistry I course at a community college, Walker and Sampson (2013) used a technique called Argument-Driven Inquiry (ADI). Used primarily in laboratory science classrooms, ADI calls for students to complete an investigation by identifying the research question, developing a method and then using it to collect and analyze data, developing a tentative argument and then refining it through discussion with classmates before submitting a final argument for grading. Seven activities throughout the semester were completed using ADI. The authors examined the written arguments generated, videotaped oral argumentation during the activities, and individual performance tasks that allowed students to demonstrate reasoning skills and content understanding. Conclusions included that students' content knowledge was improved through the process of ADI, and both written and oral argumentation skills improved with practice. The authors point out that the results of this study suggest, at least in science, that argumentation helps students better understand the course content (arguing to learn). However, students also need to learn to create arguments in the process (learning to argue). This study supports the hypothesis that the two purposes are complementary and can be accomplished at the same time.

The studies described above indicate that scaffolding of argumentation skills can be successful. However, not enough research has been conducted that demonstrates how scaffolding techniques may be used other than graphic organizers that prompt students to separate parts of their arguments.

3.2 SCAFFOLDING STATISTICAL ARGUMENTATION

Statistics education researchers have studied ways to facilitate aspects of statistical argumentation. Some of these studies, which are mainly teaching experiments, are described in this section.

Informal inferential reasoning (IIR) is a common focus in statistics education research in recent years. Pfannkuch (2006) defined IIR as "the drawing of conclusions from data that is based mainly on looking at, comparing, and reasoning from distributions of data" (p. 1). In IIR, conclusions are applied to a population based on examining a sample, either via graphical representations or numerical summary measures. Problems are often, but not always, presented in a meaningful context. The inference drawn as a result of the process of IIR is called an informal statistical inference (ISI). IIR is a key component of the thought process employed during statistical argumentation. In Pfannkuch's study, students at the secondary level were asked to compare boxplots and explain how they supported a particular hypothesis. Analysis of teacher and student communication showed the students' reasoning to be a subset of the reasoning expressed by the teacher, indicating the importance of the method of instruction. Students' conceptions of boxplots, especially the way boxplots relate to quartiles, also impacted their IIR; students with deeper understanding of boxplots tended to go beyond making basic comparisons to assess the magnitude of the difference between groups. Pfannkuch did not refer to the students' resulting explanations as statistical arguments, but they appear to fit the definition: students were asked to use the results of data analysis to support a particular hypothesis which was situated in a context.

Other studies have shown characteristics of tasks and learning environments that facilitate or are facilitated by IIR. In a study with in-service postsecondary mathematics teachers, Madden (2011) found tasks that are statistically, contextually, and/or technologically provocative, i.e., novel or controversial, tended to result in deeper levels of IIR. Pfannkuch, Forbes, Harraway, Budgett, & Wild (2013) demonstrated that using statistical computing software can facilitate IIR by quickly generating visual representations of data, bypassing onerous calculations that can derail students' thought processes. Watson (2002) showed that introducing cognitive conflict by presenting students with differing arguments yielded deeper levels of IIR.

The role of context in IIR has also been investigated. Langrall, Nisbet, Mooney, Jansem (2011) gave middle school students a task to compare two datasets in a context related to sports or popular culture. They arranged the students into six groups of three so that half the groups included a student with particular knowledge of the context, called an expert, and half did not. Each group discussed two scenarios: one in which three of the groups had a context expert and a common scenario in which none of the groups had particular knowledge of the context. Langrall *et al*. determined that students with context expertise tended to use their knowledge to bring insight to the task and justify claims. The context experts occasionally made observations that were not relevant to the task, but were not distracted by their knowledge of the scenario.

Pfannkuch (2011) observed how a group of tenth grade students interacted with context during teacher-guided IIR. Pfannkuch found that students often referred to context to make meaning of the results of data analysis. However, some students also got distracted by context, engaging in speculation about explanations for the results of the analysis or basing conclusions on their prior knowledge of the scenario rather than on the statistical evidence.

The few studies in the literature that have focused directly on statistical argumentation have not been consistent in their definition of what constitutes a statistical argument. McClain and Cobb (2001) examined the emergence of classroom norms of what they refer to as “mathematical argument[ation] in the context of data analysis” (p. 103) or “data-based arguments” (p. 104). The authors conducted a teaching experiment with a class of seventh grade students consisting of an investigation in which students generated data, analyzed it using computer software, and presented their results. McClain and Cobb studied the students' discussions during the investigation and documented the interactions leading to sociomathematical norm of what is considered to be an acceptable data-based argument. The authors do not specify what counts as an argument, but their analysis seems to indicate they mean any explanation of an individual's reasoning regarding the context, observations about features of the data, and conclusions about the data.

Only three studies have been found that describe efforts to scaffold statistical argumentation. The first took place at the postsecondary level, embedded in an enrichment course for pre-service teachers as part a scholarship program rather than in a statistics course. Derry, Levin, Osana, Jones, & Peterson (2000) designed the enrichment course in part to promote the construction and evaluation of statistical arguments. A key activity in the course was a project in which students were tasked with determining if racial bias existed in hiring practices of a company. Groups of students were given candies in an envelope, where different candy types represented races of individuals who had been hired over a five-year period. Students devised and carried out experiments, graphed sampling distributions of the data they collected, and presented their conclusions to the class. Derry *et al*. interviewed the students at the beginning and the end of the course to determine if they grew in their ability to evaluate the strength of statistical arguments presented in the media. In a qualitative analysis of the interviews, researchers determined that after the course, students more often based their arguments on statistical concepts instead of prior knowledge of the context. This growth was observed for most statistical content, though students still avoided discussing some of the more difficult statistical concepts.

Osana, Leath, & Thompson (2004) related an effort to foster in seventh grade students the development of arguments based on sampling rather than the personal opinions of one or two people. Students completed a project in which they were given reports from newspaper clippings and web pages describing a simulated scenario. The context, which students were allowed to choose, related to whether there are gender differences in self-esteem. Students were asked to use the evidence provided to present arguments for and against a new educational policy being considered by a local school board. Paper-and-pencil tests before and after the investigation showed that students improved in their ability to base their arguments on survey data rather than personal opinion.

Also working with middle school students, Hudson (2010) examined the oral statistical arguments resulting from a statistics unit in a Project-Based Learning (PBL) environment. In PBL, students learn course concepts through the use of real-life scenarios with guiding questions. Analyzing pre- and post-tests as well as videotaped daily interactions of the instructor and a focus group in the class, Hudson concluded that in the absence of proper scaffolding, middle school students do not engage in deep levels of argumentation, their arguments focus on individual data points rather than overall trends, and conclusions are based in everyday knowledge of the context rather than the results of data analysis. Hudson further discovered that teachers can facilitate better statistical argumentation by establishing clear expectations, including feedback, for discourse. Balancing content with context is also important and building on students' intuitive understandings of statistical concepts. Hudson points to a need for two kinds of scaffolding. The first is contextual scaffolding, in which the teacher helps students understand the problem and how the ideas in one context can be applied to a problem in another context. The second is social scaffolding, in which the teacher helps students learn to engage in statistical argumentation.

Statistical argumentation is still in its infancy. A key component of statistical argumentation, IIR, has been studied extensively. Research on IIR offers strategies that may be effective in scaffolding statistical argumentation, but thus far in the literature the heuristics of statistical argumentation have not been characterized. Additionally, no study has been located that attempts to scaffold statistical argumentation in the postsecondary introductory statistics curriculum. The present study will attempt to begin to fill these gaps in the literature. In this study, a teaching experiment was conducted in which postsecondary students took an introductory statistics course that included formal and informal instruction in statistical argumentation, and a series of tasks designed to support the scaffolding of statistical argumentation. Their work was then analyzed qualitatively to determine how their statistical arguments changed over the course of a semester.

CHAPTER 4: METHODOLOGY

In this chapter, methodology for the study is presented. Since this dissertation is a teaching experiment, the theory and methodology of conducting such studies are explained. Second, details about selection of participants are presented. Next, the sequence of teaching episodes, tasks to scaffold statistical argumentation, and accompanying interview protocols are described in depth. Finally, a plan for data analysis is described.

4.1 TEACHING EXPERIMENTS

The proposed study follows the paradigm of a teaching experiment, a qualitative methodology in which a teacher-researcher plans and carries out a set of instructional activities with the goal of developing an understanding of students' emerging conceptions of a given idea or task. The instructional activities, called "teaching episodes" by Steffe and Thompson (2000), take place over a period of time, often in the form of small group or one-on-one interviews focused on a task. Throughout the experiment, the teacher-researcher engages in a constant feedback loop of generating hypotheses about students' thought processes and then modifying the activities both to test hypotheses and to encourage students to refine their understanding (Steffe and Thompson, 2000).

The teaching experiment, a methodology specific to the field of mathematics education research, first emerged in the United States in the early 1970s as a combination of Piaget’s and Vygotsky’s theory. Loosely based on Piaget’s clinical interview, it was developed because existing methodologies did not account for the ongoing and iterative nature of the interactions between teacher and student, nor did they adequately explain how students make sense of mathematical ideas. Teaching experiments close these gaps by allowing the teacher-researcher to be an active participant in exploring students’ thought processes rather than simply observing them. Additionally, incorporating the interaction over time between student and teacher-researcher into the study grounds it in the way teaching and learning actually occur in real-life classrooms, meaning the results of the study are more likely to have meaningful application for practice (Steffe and Thompson, 2000).

Teaching experiments are usually comprised of a sequence of teaching episodes with the same set of students over a period of a few days or weeks. Each teaching episode is focused on an instructional activity or tool the teacher-researcher has carefully planned in advance. The teacher-researcher conducts a semi-structured interview, either while the students are completing the activity or a short time later. The interview consists of questions that gently probe students’ understanding while encouraging them to make progress. Throughout the process, the teacher-researcher engages in a continual cycle of creating and testing hypotheses about students’ reasoning. During the teaching episode, all instructional activities and interviews are captured on audio or video recording (Steffe and Thompson, 2000).

Data analysis occurs constantly during a teaching experiment. While interacting with students, the teacher-researcher analyzes their behaviors, testing against a hypothesized model. This informal analysis shapes the decisions the teacher-researcher makes in the remainder of the teaching episode. Between teaching episodes, the teacher-researcher retrospectively analyzes all available data and uses the results to plan the next teaching episode. This feedback loop continues until the teaching experiment is complete, at which time a final retrospective analysis is conducted to construct a model of each student's learning trajectory (Steffe and Thompson, 2000).

Data in a teaching experiment are examined using discourse analysis, a broad term used to describe qualitative analysis of communication. In the case of teaching experiments, the analysis is based on students' written and spoken words, students' performance on tasks, and students' cognitive behavior both during the teaching episode and over time. Toward building a model of students' developing schemas, the teacher-researcher examines the data using the theoretical framework of the study as a lens (Steffe and Thompson, 2000).

In this teaching experiment, formal and informal instruction in statistical argumentation takes place in the classroom, and students complete a series of tasks designed to support the scaffolding of statistical argumentation. Tasks include written statistical arguments and guided reflections. The students' statistical arguments and guided reflections will then analyzed qualitatively to determine how their statistical arguments change over the course of a semester.

4.2 DESIGN

4.2.1. CONTEXT OF THE STUDY

This study took place in three sections of an introductory statistics course at a small university. The overall goals of the course are for students to become proficient critical thinkers about studies involving quantitative data, and novice producers of studies involving quantitative data. Course content includes basic probability, descriptive statistics, and inferential statistics. Much of the course is focused on statistical procedures and concepts—what they mean, what information can be gleaned from them, how they are used, and in what circumstances they are used. Probability is included as background for sampling and for inferential methods. Though calculations are necessarily a part of the course, many of the calculations are completed either using the program Statistical Package for the Social Sciences (SPSS) or built-in formulas on a graphing calculator.

The three sections of the introductory statistics course met twice per week. The teacher-researcher, who holds a Master of Science degree in mathematical sciences and who at the time of data collection was a full-time lecturer at the institution for nine years, was the course instructor for all three sections. Each section was taught using the same teaching methodology, including the same instruction in statistical argumentation. As a required element of the courses, all students completed a set of tasks throughout the semester designed to scaffold statistical argumentation. As part of the assignments, every student completed four reflections, with the choice of completing them by answering a set of written questions or via one-on-one conversation outside of class time with the teacher-researcher in her office. Care was taken so that all students received the same benefit from the instruction, regardless of which type of reflection they chose and, for those who completed interviews, regardless of whether they consented to being recorded as part of data collection for this study.

Students were told verbally in class that the teacher-researcher was conducting a study and that if they chose to complete the interviews via one-on-one conversation, they could choose to have their interviews recorded for the study, but that there would be no penalty or reward regardless of their choice. It was briefly explained to the students that the purpose of the study was to examine how their writing in the tasks changed over the course of the semester, particularly in the areas of making inferences, linking to context, and stating results. Students signed up for reflections using Calendly, a web-based appointment scheduling program that allows users to reserve available times with a program administrator, in this case the teacher-researcher, who is the only person able to view identities of the appointment-holders. When students arrived for reflections, they were given the option to participate in the study. It was reiterated to the students that there was no reward if they chose to participate nor a penalty if they chose not to participate. Those who agreed signed consent forms and their reflections were audio recorded. These reflections became the interviews in the study.

4.2.2 PARTICIPANTS

The three sections of the introductory statistics course taught as part of this study together contained 75 students at the outset of the semester and 69 students at the end of the semester. The students who chose to complete their reflections via one-on-one conversation outside of class time with the teacher-researcher in her office and who signed the consent form to allow their conversations to be audio recorded were considered participants in the study. Other than signing the consent form and having the conversations recorded, the participants in the study had the same experience with the course and the same interactions with the teacher-researcher as the other students.

Over the course of the semester, all 69 students who remained in the course at the end of the semester completed the required tasks. Of those, 27 students completed one or more reflections via one-on-one conversation with the teacher-researcher, and all 27 students consented to having their interviews recorded. Thirteen students completed all interviews and assignments.

A preliminary analysis was conducted of the 13 students who completed all four interviews and assignments. Three students were chosen for a more in-depth analysis. For purposes of this study they are given pseudonyms Amber, Leah, and Kurt. These three students were chosen because their statistical arguments exhibited the widest variety of characteristics in the three criteria of statistical argumentation. Two of the students, Amber and Leah, were enrolled in the same section of the course, while Kurt was enrolled in a different section.

4.2.3 OVERVIEW OF TASKS

In this study, there were four teaching episodes. Each teaching episode included the construction of three statistical arguments—first by the instructor incorporating suggestions from the class, second by groups of two to three students working together during class time, and third by each student working independently outside of class. Following the third assignment of each teaching episode, each student completed a reflection; students were allowed to choose either a written reflection or a verbal reflection. The four teaching episodes took place at three- to five-week intervals during the semester. They were scheduled to coincide with 1) the coverage of sampling; 2) data description methods; 3) one-variable, one-sample inferential methods; and 4) two-variable or two-sample inferential methods. Statistical arguments in each teaching episode were based on the statistical methods and constructs in the curriculum up to that point in the semester; thus, they increased in complexity over time.

The format of each of the tasks was the same. Students were given a document that provided: 1) a research question that requires a statistical argument to answer, 2) sampling and data collection procedures, 3) results of a data analysis in the form of output generated using SPSS, and 4) instructions to construct a written argument to answer the research question. All data for these tasks were simulated in order to make it possible for key concepts to be illustrated and contexts to be familiar. Students submitted all work using the school's learning management system, a Moodle-based program called MyCourses.

4.2.4 RATIONALE FOR THE TEACHING EXPERIMENT

In this teaching experiment, scaffolding occurred in three ways: 1) structure of the assignments, 2) one-on-one interaction between students and the teacher-researcher, and 3) the cultivation of a classroom culture to support statistical argumentation. The ways in which statistical argumentation was scaffolded are key in explaining the rationale for the elements of this teaching experiment. The second part of the rationale is background information about the three pilot studies, which were prior semesters of conducting this program of statistical argumentation, and the changes it underwent to make it more successful in the fourth semester, the one in which this teaching experiment occurred.

In this section, the teaching experiment is described in-depth. The elements are organized according to the type of scaffolding employed. Rationale is explained in terms of how elements scaffolded statistical argumentation and how elements evolved over the three pilot studies.

4.2.4.1 STRUCTURE OF THE ASSIGNMENTS

The first form of scaffolding employed in this teaching experiment is in the structure of the assignments. Statistical argumentation is scaffolded first horizontally in the construction of each teaching episode and second, vertically in the sequence of the four teaching episodes. Description of both kinds of scaffolding follow, and serve to provide more detail about this teaching experiment.

4.2.4.1.1 HORIZONTAL STRUCTURE

Each teaching episode was structured following four phases. In the first phase, the teacher-researcher presented an example to the class (labeled sample arguments 1, 2, 3, and 4), the first task of each teaching episode. She led a discussion of the information provided and guided students through an oral statistical argument of the results. She showed students a written statistical argument for the example based on the model of statistical argumentation adapted from Abelson (1995), which was described in section 2.3.2 of this dissertation. They were allowed to read it but not keep it. They were also given a list of elements to include in their arguments (included in Appendix A). The model was presented with the three broad criteria (linking to context, articulating the results, and making inferences) with the addition of a fourth criteria, synthesizing. For each criteria, specific—and therefore clearer to the students—elements to include were presented for each of their statistical arguments. To aid student understanding of the factors to consider, the teacher-researcher pointed out to the students how the factors were incorporated into the sample argument.

Finding the appropriate amount of guidance for students proved to be the most challenging aspects of developing this teaching experiment, and underwent more adjustment throughout the three semesters prior to the teaching experiment. The first semester it was implemented, students were given only the sample arguments in class and not allowed to keep them; they were not given a list of elements to consider. Students seemed to need more guidance and clarity of expectations because the content included in their arguments varied widely, as did the quality of the arguments. The second semester, the example arguments were posted on the course site for students to access. The result was that students strongly tended to parrot the examples rather than write in their own words, at times even modifying the example arguments in such minor ways that they did not make sense for the arguments they were being asked to construct. The third semester, students were given the example arguments only in class again, but with accompanying lists of elements to include in the arguments they were being asked to construct. The combination of providing a list of elements to include and presenting a sample argument but not allowing students access to it while they constructed their own arguments appeared to give the right amount of guidance. In response to a small number of students who wrote their arguments as bullet points in the third semester of implementing this set of assignment, one adjustment for the semester in which this teaching experiment occurred was to include an instruction on the list of elements that arguments should be written in paragraph form, not bulleted points. Only one other adjustment was made in the last semester; handouts were created for the last two assignments to provide written instruction to supplement the in-class instruction on how to read SPSS output for the one-variable t-test and the independent samples t-test. The handouts were created to compensate for the fact that the last part of the course felt a bit rushed due to the two class days lost for inclement weather.

The second phase of each teaching episode took place in the same class period as the first. Immediately following the example argument and associated whole-class discussion, students were assigned groups of two to three to complete the second task of the episode, (labeled 1a, 2a, 3a, and 4a) generating a statistical argument based on a different research question, sampling and data collection procedure, and results. In prior implementations of these assignments, it was clear that when groups were assigned rather than self-selected, students were more productive in class and generated better arguments overall. Groups were randomly assigned using a random number generator, though the teacher-researcher made minor adjustments to the groups to promote balance of student strengths or to adjust for student absences. While students worked in class, the teacher-researcher remained at the front of the room but near enough to hear students' discussions. She was available to answer questions and occasionally offered guidance, but tried to keep her involvement at a minimum so students could work as independently as possible. Students worked twenty to thirty minutes in class and were then instructed to complete their work outside of class. Feedback, which will be discussed further in section 4.2.4.2, was provided on group assignments before the third phase of the teaching experiment.

In the third phase of each teaching episode, students completed the third task of the teaching episode (labeled 1b, 2b, 3b, and 4b). For this task, each student worked independently outside of class to create a statistical argument based on a different research question, sampling and data collection procedure, and results. To ensure arguments reflected each student's thoughts and words, students were given instructions that they were not to consult with anyone except the teacher-researcher, and they were asked to sign a pledge (which is standard at the university where this study is conducted) when they submitted the assignment that they had not discussed it with anyone besides their instructor.

The fourth and final phase of each teaching episode was a reflection. Students were allowed to choose either a written reflection or a verbal reflection. Both methods of reflection focused students' attention on features of their individual arguments related to the statistical language used in the argument as well as the criteria of context, articulation of results, and inference. Reflections will be discussed further in section 4.2.4.2.

The steps in each teaching episode as described in the preceding paragraphs provide horizontal scaffolding of statistical argumentation. To summarize, the first argument was constructed by the class as a whole with strong guidance from the teacher-researcher. The second argument was constructed with minimal help from the teacher-researcher; instead, by working in small groups, students had guidance from other students, refined their thinking through discussion with group members, and combined their collective ideas into an argument. After feedback was given on the group arguments, the third argument was constructed independently, with no support from the teacher-researcher or classmates. This withdrawal of support as students are able to complete tasks independently is the essence of scaffolding.

4.2.4.1.2 VERTICAL STRUCTURE

The second type of scaffolding in this teaching experiment was vertical. During the semester, there were four teaching episodes, which progressively included more content as students were introduced to new concepts in the course. The following paragraphs describe how each teaching episode fit into the introductory statistics curriculum, the course concepts upon which it drew, and how the criteria for statistical argumentation were applied.

Table 1. Course schedule for Introductory Statistics

Week	Topic(s)
1	Introduction to the Language of Statistics Introduction to distribution, center, and spread
2	Discrete Probability Distributions
3	The Normal Distribution
4	The Central Limit Theorem Sampling, Critical Thinking about Statistics
5	**Teaching Episode #1: Sampling** Graphical Statistical Methods
6	Measures of Center and Spread
7	Measures of Position
8	Mid-term Exam Review and Mid-term Exam

9	**Teaching Episode #2: Descriptive Statistics** Confidence Intervals for the Mean
10	Confidence Intervals for a Proportion Intro to Hypothesis Testing and P-values, the Z Test
11	The T Test for a Mean **Teaching Episode #3: Inferences on One Variable**
12	Z test for a Proportion Chi-Square Test for Independence
13	Testing the Difference between Two Means **Teaching Episode #4: Inferences on Two Samples**
14	Choosing which formula to use Correlation and Linear Regression
15	Final Exam Review and Final Exam

The first teaching episode took place after a lesson on sampling techniques and critical thinking about statistics. This teaching episode focused on the context and inference criteria for statistical arguments. Students were asked to consider whether the research question and data collection procedures made sense for the scenario being discussed, the context. Additionally, students determined what conclusions or inferences made were warranted based on the sampling procedure and data collected, the inference criteria. Finally, they evaluated the credibility of the results, linking to context. At this stage, articulation was limited to simple percentages since statistical methods had not been covered yet in depth.

The second teaching episode occurred four weeks after the first. This teaching episode was planned to occur two weeks after the first, but two class days missed due to inclement weather forced a delay that included a mid-semester break. The schedule change also forced the second, third, and fourth teaching episodes to occur more closely together than originally planned. By this point in the course, students had learned descriptive statistics. This unit consisted of graphical techniques of displaying univariate data, including histograms, bar charts, pie charts, and boxplots. It also contained numeric summary measures for samples, including measures of center, spread, and position. Descriptive statistics are part of the articulation criteria of statistical argumentation, so articulation was the primary focus in the second teaching episode. However, context and inference were still be important in synthesizing all the information given to create an argument to answer the research question they were given.

The third teaching episode took place two weeks after the second one. At the time of this teaching episode, students had learned about one-sample confidence intervals and hypothesis tests. The three scenarios presented in this teaching episode were all be based on the one-sample t test for the mean. Keeping the arguments based on the same kind of hypothesis test minimized the confusion that may have resulted from asking students to read output from different SPSS procedures. The t test was chosen because it is a common hypothesis test and also because of the availability of a convenient measure of effect size, Cohen's *d*. Discussing hypothesis tests, including effect size, is part of articulating the results. The third teaching episode focused primarily on articulating results and making inferences, though context was interwoven throughout.

Two weeks later, when the fourth and final teaching episode took place, two-variable and two-sample hypothesis tests had been covered. Topics included correlation, regression, independent samples t tests, paired t tests, and two-sample proportion tests.

For the same reasons the one-sample t test was chosen for all three arguments in the third teaching episode, the independent samples t test was chosen for all three arguments in the fourth teaching episode. Articulation and inference criteria were the primary focus on this task, although context was still important throughout the argument.

As described in the preceding paragraphs, the teaching episodes in this teaching experiment were designed to provide vertical scaffolding. The articulation of results in the tasks of the first teaching episode required an understanding only of basic percentages, allowing students to focus on the context and inference aspects of their arguments. In the second teaching episode, concepts of distribution, center, and spread were incorporated. In the third and fourth teaching episodes, the addition of hypothesis tests and effect sizes provided more advanced statistical content to the arguments—and thus more complex arguments. Statistical argumentation was scaffolded in the way these teaching episodes prepared students for progressively more advanced statistical arguments.

4.2.4.2 ONE-ON-ONE INTERACTION

The next type of scaffolding in this teaching experiment occurred via one-on-one interaction between students and the teacher-researcher. One-on-one interaction took place anytime the teacher-researcher offered either written or verbal feedback. It also took place during the reflections, particularly the verbal reflections which allowed for back and forth discussion.

4.2.4.2.1 WRITTEN FEEDBACK ON ASSIGNMENTS

Feedback on assignments was a crucial element of scaffolding statistical argumentation. Feedback took the form of written comments delivered through the learning management system. Efforts were made consistently through the semester to ensure students had feedback on each assignment before starting work on the next. In order to give prompt feedback on all eight assignments to such a large number of students, it needed to be brief. As a result, comments tended to explain the reasons points had been deducted from students' grades, make suggestions for improving future arguments, and praise excellent work. Examples of feedback are:

- "Excellent!"
- "Very good job. I would like to see more clarification in how the sample is biased and how you think the answers may be skewed (presumably cheating would be underreported, but I want to see it spelled out)."
- "The first paragraph is excellent. However, you didn't state all the means, which are probably the most important part of this argument. Also, a low standard deviation doesn't mean there isn't any variation, just less variation than the other groups."
- "The numbers on the boxplot next to the outliers are observation numbers, not data values.
  This test is a t-test, not a z-test, so the test statistic should be stated as t.
  Effect size is always positive. Need to take the absolute value.
  It would help to have more discussion about the hypothesis test, link it to the effect size and to the scenario."

There was one exception to written feedback on the assignments. For students who completed reflections verbally, feedback on the individual assignments was given at the conclusion of the interviews. The reason is, it was hoped in the interviews that students would be guided to improve their own arguments. Providing written feedback in advance would compromise the process of scaffolding in the interviews.

The other part of providing feedback on student work was assigning grades. Though the process of statistical argumentation was of greater concern to the teacher-researcher, grades were nevertheless necessary and important to the teaching experiment. They were necessary for the practical reasons of course management and student accountability, but they also served as a way of communicating to students the quality of their work.

Grading was based on a set of rubrics, one for each assignment. Each rubric contained three items: 1) linking to context, 2) articulating results, and 3) making inferences. The rubric assigned ten points for linking context and ten points for making inferences. In the first argument, ten points were assigned to articulating results, in the second argument 20 points were assigned for articulating results, and in the third and fourth arguments 30 points were assigned for articulating results. The sum of the points on each rubric determined the weights for each assignment. Weights for the reflections were chosen to be high enough to motivate students to complete them but low enough to acknowledge that the reflections required less work than the arguments. Table 2 shows the weights that were assigned to each task. During development of these assignments, the first attempt of making rubrics available to the students resulted in confusion; subsequently, the rubrics were used solely for the teacher-researcher as a guide for allocating grades. The statistical argumentation average, which counted for 25% of the final course grade, was calculated as the sum of the points earned divided by 400, the maximum number of possible points.

Table 2. Weights assigned to tasks

Teaching Episode	Small group argument	Individual argument	Reflection
1	30	30	15
2	40	40	15
3	50	50	15
4	50	50	15

Feedback is a crucial element of scaffolding in any teaching and learning situation. In this teaching experiment, feedback was offered via comments and grades. Providing both a measure of the quality of their arguments and specific comments about their arguments was a way of guiding students to improve their future arguments.

4.2.4.2.2 INTERVIEWS

The next type of one-on-one interaction between students and the teacher-researcher occurred during the reflections, both written and verbal. While there was value to the written reflections for students who chose them, the focus in this section is on the verbal reflections, since they were monitored in this study. Interviews occurred at the end of each teaching episode, lasted between 10 and 15 minutes, and focused on the individual arguments students had recently submitted.

Each verbal reflection was a guided but somewhat loose structure interview designed to guide students through the individual argument task for the current teaching episode. Though the interview was focused on the most recent task, it was based on a review of all available data for the student up to that point. At the outset of the interview, the student signed a consent form (which included four signature lines, one for each interview) and was reminded that the conversation was solely for the purposes of understanding their thinking and would not be used to evaluate or grade the individual arguments. The interview then progressed to a discussion of the elements of the individual argument. Throughout the interview, questions and indirect guidance were designed to encourage the student to:

- incorporate the context of the study into the statistical argument;
- compare, contrast, and combine elements of the results of the statistical analysis to develop a holistic understanding of the results;
- better articulate the results of statistical analysis into the statistical argument;
- make inferences and incorporate them into the statistical argument;
- synthesize context, articulation, and inference toward creating a cohesive statistical argument;
- clarify and refine the use of statistical language; and
- clarify and refine the conceptions about statistical structures and processes.

Not every interview included all of the above elements, and some required additional elements. In some instances, it became necessary to pause and conduct some direct instruction on statistical concepts or procedures before returning to the interview. The working principle was that the teacher-researcher conducted the interview with the goals of 1) helping students further develop their statistical argumentation strategies as well as their understanding of statistical structures and processes, and 2) building a model of students' statistical argumentation.

4.2.4.3 CLASSROOM CULTURE TO SUPPORT STATISTICAL ARGUMENTATION

The final method of scaffolding statistical argumentation is through the cultivation of a classroom culture to support statistical argumentation. It is imperative that the work of preparing students to construct effective statistical arguments occur throughout the course and not just during the teaching episodes. Accordingly, the curriculum was presented with ideas of context, articulation of results, and inference in mind. These were accomplished through instruction in critical thinking, interpretation of the results of statistical calculations, and continuing emphasis on the use of statistical language. This approach served not only to prepare students for statistical argument tasks, but more importantly, they focused the course on the statistical literacy, reasoning, and thinking skills described in section 2.1. The following paragraphs describe some of the efforts of cultivating a classroom culture to aid in the scaffolding of context, articulation of results, and inference in statistical argumentation.

4.2.4.3.1 CLASSROOM CULTURE OF LINKING TO CONTEXT

During the semester, one lesson about critical thinking in statistics followed by ongoing prompts about the language of statistics were geared toward scaffolding the first criteria, context. The lesson was in week 4 and focused on critical thinking about statistics. Students were given examples of studies, some flawed, and guided through evaluation of 1) whether sampling and data collection methods made sense for the scenario presented, and 2) whether the results of the study were credible in the scenario presented. Two of the examples presented were real-life scenarios of studies similar to the ones given in the first teaching episode. The first example was that of Shere Hite's (1989) study Women in Love, in which surveys were sent only to women's groups, and the response rate was 4.5%; one result of the study was that 70% of the married women in the sample were having sex outside their marriages. Another example was a 1993 poll conducted by the Roper organization in which respondents were asked, "Does it seem possible, or does it seem impossible to you that the Nazi extermination of the Jews never happened?"; the result of the study was that 22% of the sample indicated it was possible the holocaust had never occurred (Ladd, 1994). In both of these scenarios, students were asked if the results aligned with their expectations; i.e., they were asked to compare the results of the study against their general knowledge of the contexts. Most said the results of both studies did not seem credible based on their general knowledge, and the students did not trust the studies as a result. The sampling problems, which will be discussed under inference, contributed to their skepticism.

Another way context was incorporated into the classroom culture was through consistent use of it in the language used throughout the course. Except in rare circumstances, examples and homework problems were presented in context, some real-life scenarios and some simulated. With the examples, the teacher-researcher tried to link to context in three ways. First, students were asked why the study might be valuable to conduct, or how the results might be used. The question is closely related to Abelson's concept of interestingness, which he considers to be a part of a good statistical argument, but also serves as a reminder of the context of the study. Second, during lectures when discussing the results of calculations, the teacher-researcher made a point of prompting students to answer what the results meant in the context of each problem. Third, for example problems in which the contexts were familiar, students were asked to evaluate the reasonableness of the results. Consistently using these three strategies created a classroom culture in which datasets were approached with the context in mind.

4.2.4.3.2 CLASSROOM CULTURE OF ARTICULATING RESULTS

Much of the curriculum in the introductory statistics course focuses on the second criteria, articulation of results. Coverage of these topics in the curriculum is necessary to scaffold statistical argumentation, but it is not sufficient. Toward preparing students to use these topics in their statistical arguments, two particular efforts were made: 1) organizing the material around concepts of distribution, center, and spread; and 2) modeling language of interpreting these concepts.

One step toward preparing students to articulate results in their statistical arguments was framing the course around the concepts of distribution, center, and spread. This was a substantial adjustment in the organization of the course but required only one day of instruction to introduce the concepts, followed by brief introductions during lecture tying the material back to them. The lesson introducing distribution, center, and spread defined the three concepts; demonstrated how they interrelate; and gave an overview of them for samples as well as populations, in both discrete and continuous scenarios. In the introductory lecture, measures of center and spread were introduced in concept but calculations were reserved for later in the course. In particular, standard deviation is a more difficult topic, so in the introduction it was explained simply as “typical distance” (the teacher-researcher took care to distinguish it from average deviation). Students practiced describing distribution shapes, identifying the population associated with a given sample, choosing appropriate measures for center and spread, and classifying measures as parameters or statistics. This approach of framing the course around distribution, center, and spread was valuable in scaffolding statistical argumentation; however, it also served as a way to highlight the similarities and differences of how sample and population are described, and provided a “big picture” to keep in mind throughout the course, which is especially useful for calculations that can be tedious.

The second way of scaffolding articulation of results in the classroom culture was modeling it during lectures as course concepts were covered. One case of modeling articulation of results came at the end of the lecture on measures of center and spread. Students were given an example of hypothetical means and standard deviations of prices for three brands of aspirin. The teacher-researcher guided students through discussing the values, as in:

> Brand B is on average the most expensive brand, followed by Brand C and then Brand A. Because Brand A has the lowest standard deviation of the three brands, the prices vary less than the others. Since it has the lowest price with the highest degree of consistency, Brand A is the best choice when it comes to price.

Another example was in the lesson on measures of position. Students were coached to consider whether any outliers in a dataset were within the range of possible values of the data; e.g., if an outlier of 30 were present in a dataset of how much television people watch per day, it must be an error since the range of possible values is 0 to 24, inclusive.

Finally, when effect size was covered, the teacher-researcher discussed with students how to interpret it in conjunction with the results of the associated hypothesis test. Two examples are:

- Because the p-value is less than .05, the mean salary of public school teachers in North Carolina can be concluded to be lower than the nationwide mean. At .52, the effect size indicates that the difference is substantial but not huge.
- The two-sample t-test yielded a p-value of .17, so we fail to reject the null hypothesis and cannot conclude that the mean number of calories burned by walking briskly for 30 minutes is less than the mean number of calories burned by jogging for 30 minutes. An effect size of .09 further supports that even if there were a statistically significant difference, it would be very small.

4.2.4.3.3 CLASSROOM CULTURE OF MAKING INFERENCES

Finally, there was a strong and repeated emphasis throughout the course on the concept of statistical inference. This emphasis began the first day of class, in a lesson introducing key terms and concepts. The lesson included definitions of sample, statistic, population, and parameter, among other ideas. Statistical inference was presented as a primary goal of statistics. In that lesson, informal statistical inference was introduced by presenting a simple scenario, a sample of hammerhead sharks having a mean length of 12 feet, and guiding students to the simple inference that the population mean would be expected to be approximately 12 feet. This type of informal statistical inference was made frequently in examples given in class lectures throughout the descriptive statistics unit.

Another key piece of creating a classroom culture to support scaffolding of making inferences was a lesson that occurred in the lesson in week 4 on sampling and critical thinking in statistics. The lesson included a discussion of sources of bias in sampling and how inferences may be limited as a result of sampling bias. It also included a discussion of common mistakes in interpretation of statistics, particularly inferring causality from observational studies, and making conclusions on standalone statistics that are stated without a basis for comparison. An example of sampling bias, Hite's study Women in Love, previously discussed in section 4.2.4.3.1 of this dissertation, was presented to students; they identified bias both in sending surveys only to women's groups and in the low response rate, resulting in statistics that can be inferred only to a very limited population. Students also read about an example of a problematic data collection procedure, the Roper poll about the Holocaust, also discussed in section 4.2.4.3.1; they identified a confusing question as the likely cause of an inflated proportion of people who doubted whether the Holocaust had occurred. These same kinds of questions arose in example problems throughout the semester, encouraging students to look at summary statistics or results of hypothesis tests with a critical eye.

A classroom culture strong in linking to context, articulating results, and making inferences was a necessary step in scaffolding statistical argumentation. The instruction in these norms did not cause the course to become entirely focused on statistical argumentation. All the same concepts were included in the curriculum as in previous semesters. The cultivation of a classroom culture to support statistical argumentation required only some rearranging of the material and a conscious effort to be consistent in linking to context, articulating results, and making inferences. Additionally, these steps served the larger purposes of critical thinking, application of statistical concepts, and use of statistical language.

4.3 DATA COLLECTION

Approval from the Institutional Review Board (IRB) was obtained before any data was collected. Written data was collected throughout the semester from students' written work on the eight tasks they completed during the four teaching episodes. Additionally, audio recordings were made and transcribed of one-on-one semi-structured interviews.

4.4 DATA ANALYSIS

Data analysis occurred throughout the study, both formally and informally, as part of the continual feedback loop that is fundamental to teaching experiments. Though data analysis in teaching experiments is an ongoing process, generally speaking it can be categorized by the phase of the study in which it takes place: during teaching episodes, between teaching episodes, and at the end of the study. In this classroom teaching experiment, observations were not limited to the teaching episodes. As a result, the first phase of analysis occurred on a daily basis, beginning with the first day of class. All analysis was performed in Microsoft Word by the teacher-researcher in consultation with her dissertation chair.

First, on a daily basis, the teacher-researcher analyzed students' words, work, and actions. Primarily, these consisted of the students' statistical arguments and interviews. However, they also included observations from individual questions, class discussions, small-group interactions, quizzes, and tests. Any available information about students' understanding contributed to the conclusions at the end of the study. Detailed notes were made for use in the retrospective analyses.

Second, the teacher-researcher conducted a more in-depth retrospective analysis after each teaching episode, following the conclusion of the interviews. Analysis at this phase consisted of detailed examination of all prior work, interviews, and notes related to each student. The teacher-researcher observed changes over time in order to follow up on those aspects of each student's learning trajectory. Commonalities across cases were observed at this stage to help the researcher to build an overall model.

Lastly, after the teaching experiment was completed, i.e., after the end of the semester, the data was analyzed retrospectively in two phases. The first phase was a preliminary analysis of the thirteen students who completed all tasks and interviews. The purpose of this phase was to identify commonalities and differences in the learning trajectories of the students. At the conclusion of this phase, three students' work was chosen for more in-depth analysis and presentation as case studies. The students were chosen to represent the widest available variety of characteristics of statistical arguments.

The second phase of the retrospective analysis was to develop the three case studies. At this stage, all the available data was re-examined for Amber, Leah, and Kurt. A detailed analysis of the four individual statistical arguments for each of the three students was written, using interview transcripts and teacher-researcher notes as supporting data. Following the analysis of each student's work, trends were identified in the development related to the research questions, and a final model for the learning trajectory of each student was constructed. The final case studies were written to show these trends. Following the analysis of individual cases, commonalities and differences were observed across cases.

CHAPTER 5: RESULTS

In this chapter, three case studies will be presented. For each case study, the participant’s progress in the statistical argumentation criteria will be described, and then conclusions discussed.

Each student completed eight statistical arguments, consisting of four group arguments and four individual arguments. The individual statistical arguments are the primary unit of analysis. Because the group statistical arguments do not reflect the thinking of each student, they are not used as evidence of the student’s statistical argumentation. Interviews are used insofar as they help make sense of the scaffolding of context, articulation of results, and statistical inference; for example, they may be used to clarify a student’s intended meaning of a passage, to provide additional information about the student that affects his or her perception of context, or to demonstrate a time when a student gained a better understanding of a statistical concept.

5.1 CASE STUDY 1: AMBER

The subject of the first case study, Amber, is a sophomore nursing major. She is an African American student from the Southeastern United States. She is a very conscientious student, but she struggles with statistical concepts and calculations. In the early part of the course, a health issue and the resulting class days she missed impacted her performance. By the second half of the course, both her health and her performance in the course had improved. She reports that the course material did not come easily, and she required many hours of study to grasp hypothesis testing, a key concept both in the second half of the course and in the last two tasks. Amber’s work was chosen for analysis because it shows an example of statistical arguments generated when a student struggles with the underlying statistical content. It was chosen also because her learning trajectory is different from that of other students; most students’ comfort level with statistical concepts either stayed at the same level of quality or decreased slightly as the material accumulated in the course, whereas Amber’s comfort level improved.

5.1.1 LINKING TO CONTEXT

In the first task, Amber struggles to understand the context and incorporate it into her statistical argument. The research question in the first scenario is, “Do teens pay attention to local and national politics?” Amber recommends altering the research question to “Do teenagers age (13-19) watch the local news, and stay aware of the current events?,” a change she writes would improve it in two ways. First, she claims it would cause the survey to “pertain to teenagers from ages 13 to 19.” For Amber, defining teenagers as individuals from ages 13 to 19 serves to clarify the age range of the population of interest. Second, she states “the question would ask if they are staying up to date with current news,” as a benefit of her proposed research question. She is suggesting a shift in the context of the study away from politics toward news and current events. Instead of accepting the research question and trying to bring the study into alignment with it, Amber is advocating changing the context altogether. She subsequently suggests focusing on topics such as celebrities that are of more interest to teens. These recommendations indicate Amber misunderstands the expectations in the assignment, which are that students should attempt to answer the given research question rather than try to change it.

In the interview associated with the first task, Amber is asked to suppose she wants to keep the same research question, and discuss what survey question might help answer it. Amber then mentions cyberbullying. She is then asked if she sees cyberbullying as a political issue, and she replies in the affirmative. In Amber's context, a question about cyberbullying is more relevant to whether teens pay attention to politics than the given question about suffrage. It is a different specific political issue which to her addresses the more general question raised in the research question. This shows Amber is struggling to grasp the intended context presented in this assignment and the use of context in the task, since her understanding is that a survey question about a specific issue can answer a broader research question.

In the second task, Amber still struggles to understand the intended context and the expectation in the assignment to try to answer the research question that is provided. The research question given in the problem is "Is there a difference among the average speeds of drivers of neutral-colored cars, red cars, and black cars?" Amber begins her argument by recommending a revision to the research question:

> I feel as though the research question is difficult to interpret, instead the research question should be revised and state "Do drivers of certain colored vehicles such as neutral-colored, black, or red cars speed more often". This way the study is based on the vehicle color and it is easier to interpret.

One of the ways in which she claims her research question improves upon the original one is that the revision bases the study on the color of the vehicle. The given research question refers to the "drivers of neutral-colored cars, red cars, and black cars" while Amber's question refers to "drivers of certain colored vehicles such as neutral-colored, black, or red cars". This is similar to her recommendation in the first task to define teenagers as those between the ages of 13 and 19, in that it is a slight rephrasing which to her changes the meaning of the question. In both instances, she is focusing on specific aspects of the context rather than the larger issue.

A concern arising from the second task is that Amber refers to the units in the problem as “units” rather than “miles per hour”. When asked about it in the interview, she says she is confused because even though the cars in the scenario were all clocked at 60 miles per hour, the assignment did not state the units of the data. Worldwide, vehicle speeds are typically measured in kilometers per hour or miles per hour. Though she is attempting to relate to the context of the study, this missed connection reveals her thinking may not be completely grounded in the context.

In the third task, Amber approaches context in a similar way as in the previous two tasks. The research question given in the scenario is, “Does an appliance company make a profit when they offer a two-year warranty on a certain brand of refrigerator?” and it is accompanied by the additional information that “Customers pay $175 for the warranty, so if the amount of claims the company must pay to repair or replace the refrigerators is less than $175 on average, they will make a profit.” Amber evaluates the research question as needing clarification on the brand name of the refrigerator. She suggests changing it to “Does an appliance company make a profit when they offer a two-year warranty on a specific brand of refrigerator such as Frigidaire?,” saying the benefit will be that only Frigidaire products are sampled. To her, the “certain brand of refrigerator” described in the given research question is different from “a specific brand of refrigerator such as Frigidaire.” This shows she is again focusing on a detail rather than the overall context.

In the interview about the third task, questions about outliers revealed further confusion about the context, as the following exchange shows.

Amber: There was a good amount of outliers in this one. Because there was one that was right in between 500 and 600, and then from 800 up to 1000.

TR: OK. Do you think those are legitimate observations or do you think they might be errors?

Amber: I think those are legitimate. I don’t think they’re errors.

TR: What does an observation of 850 indicate?

Amber: An observation of 850…(trails off, pauses)

TR: So a value of 850 in this dataset would mean that for that refrigerator…?

Amber: 850 repairs?

TR: The units are…?

Amber: No, the units are dollars so it would be 850 dollars.

TR: Yeah

Amber: But then the warranty is only 175.

TR: The warranty *costs* 175 dollars.

Amber: So that would be the warranty and the cost of the refrigerator together?

TR: No, 175 is what the company gets. If the refrigerator breaks, then the company has to pay to replace or repair, however much that costs.

Amber: Oh, so that would be up to 850 dollars’ worth of repairs.

TR: Yeah

Amber: I think that could be legit, depending on if it’s a horrible refrigerator.

This exchange shows that engaging Amber with the raw data allows the teacher-researcher to find out if Amber’s context is the same as the intended context. In this case, since they appear not to be the same, it presents an opportunity to bring Amber’s context closer to the intended context.

In the fourth task, there is not enough information to determine whether Amber understands the context. She evaluates the research question, “Is there racial bias in sentencing for felonies?” only vaguely, saying it is “plausible, and makes logical sense.”

An attempt in the interview to determine her understanding of context by asking if she thinks there is racial bias in sentencing for felonies again yields little information; she responds, “I guess it’s kind of hard to tell. I don’t really know.” For purposes of this study, it may have been valuable to press her further on this, but ethical and pedagogical concerns about flustering her regarding a sensitive topic outweighed the value of such a line of questioning to this dissertation. If she fully understands this context, this is an indication of growth from the first three tasks. However, there is not enough information in her written argument or interview to determine whether she does or does not understand the context.

In summary, Amber’s understanding and use of context change little over the course of the teaching experiment. She tends to focus on minute details of the contexts rather than the overall scenario given in the problems. In the interviews, she expresses confusion regarding the scenarios, and it is only after discussion with the teacher-researcher that her contexts seem to be aligned with the intended contexts.

5.1.2 ARTICULATING RESULTS

This section is divided into four sub-sections. In the first section, Amber’s discussion of center and spread are presented; these are combined because in her tasks, she writes about these aspects together and in a similar way. In the second section, Amber’s characterization of distribution, especially outliers, is described. In the third section, Amber’s use of hypothesis testing is presented. In the fourth section, Amber’s use of context through articulation of results are discussed.

The first task yields little information about Amber's ability to articulate results. This is partly because the results in the first task are nothing more advanced than relative frequencies—there is no center, spread, distribution, or hypothesis test to discuss. In Amber's case, it is also uninformative because she does not articulate the results of the study other than to state the sample size. Her written argument consists of evaluating the information provided rather than attempting to answer the research question. For these reasons, the analysis in this chapter focuses on the final three tasks.

5.1.2.1 CENTER AND SPREAD

In Amber's second task, she states the means of the three groups but does not compare them, which in this scenario is fundamental to answering the research question. For spread, she reports and compares the standard deviations; though there is some value in comparing standard deviations, for this task it is not as important as comparing means. The interview reveals that Amber struggles to understand spread; when asked to characterize and compare the centers and spreads of the three groups, she seems unsure and looks at the side-by-side boxplots. Eventually she discusses the numerical measures, but as she is talking she looks at the boxplots. Though she does not articulate the connection between the graphical and numerical representations of center and spread, the attention she pays to the boxplots while discussing the numerical measures shows she is considering both. Her understanding of the connections between numerical and graphical representations of spread would have been an interesting area to explore in the interview, but the subject was not broached because it was near the end of the interview and Amber was showing signs of fatigue.

In the third task, Amber's use of center and spread change in one key way: she specifies she is choosing the mean and standard deviation as measures of center and spread. This is important because it shows she connects these measures to the three themes (center, spread, and distribution) of data analysis visited almost daily in class. She does not compare the mean to the cutoff point at which the company will make a profit, which would help answer the research question. Despite the correct use of standard deviation to characterize spread in the written argument, in the interview she still struggles to recall what measures are relevant in describing spread. When asked in the interview to characterize spread, she first mentions the skewness in the distribution, then points to the confidence interval before stating the standard deviation. When prompted to consider the range to determine how spread out the data is, she cannot recall how to find it or what it means. Amber is only able to discuss the magnitude of the spread when she is guided to consider the minimum and maximum. It appears that at this point in the semester, she does not fully understand measures of spread and how they can be used in a statistical argument, though she is able to use them in a written argument.

In the fourth task, Amber again states the centers and spreads, specifying that she is using the mean and standard deviation as measures. She does not compare values of the mean and standard deviation, which would help in addressing the research question of whether there is racial bias in sentence lengths for felonies. There is virtually no change in the way Amber discusses spread in her third and fourth written tasks, but the interview shows growth. When asked to describe the spread of the data, she immediately replies, "the spread is the standard deviation and the Caucasian was 8.22 and the non-

[Caucasian] was 7.105 and that was a little bit but it was not that much, either." By the time of the fourth task, she easily identifies the primary measure of spread as the standard deviation, states the values for both groups, and evaluates their magnitude. This is quite an improvement from the previous task, in which she could not recall which measures or graphs to use describe the spread of a dataset.

To summarize, the ways Amber describes center changes little over the course of the teaching experiment; she states the means without comparing them across groups or to a value important in the scenario. Her use of spread does not change in the written tasks, but her ability in discussing it improves, from not knowing how to begin in the second interview to quickly reporting and comparing groups in the fourth interview. In the end, though her arguments could be improved by comparing across groups, she is confident in discussing center and spread.

5.1.2.2 DISTRIBUTION

Amber's discussion of distribution is another area in which there are subtle improvements during the teaching experiment. While her incorporation of distribution shape is unremarkable, consisting in each task of the correct identification of the shape name without further elaboration, her discussion of outliers undergoes some changes.

In the second task, Amber's discussion of outliers includes an error: "By looking at the neutral-colored vehicle boxplot, I can see that it has an outlier of 2 units and the black-colored vehicle boxplot has an outlier of 47 units." The numbers 2 and 47 on the boxplots actually refer to the observation numbers of the outliers, not the values of the data points. Beyond stating the outliers, she does not use them to draw any conclusions from the data.

In the third task, Amber’s characterization of outliers is correct and she makes an attempt to discuss the outliers at greater length than she did in the second task. However, the language she uses is not clear. She writes, “The distribution is right-skewed, and there are multiple observations numbers. There is an observation of 191, 150, 126, and 60 at 500.00 to 600.00 claims paid and 44,179,200 at 800.00 to 900.00 claims paid.” This is unclear in three ways. First, she refers to the outliers as “observations”; while it is true they are observations, they are being discussed here as outliers. Second, she lists the observation numbers, which do not provide information about the outliers unless the raw data is available. Finally, she lists the ranges in which the observations lie, but as was discussed in section 5.1.1, the description of them as “claims paid” shows her confusion about the context of the data. In the interview, she correctly reads the outliers from the boxplot, this time not confusing the observation numbers for data values. Overall, her understanding of outliers is expanded and more correct than in the previous task. However, her written explanation of them needs clarification.

In the fourth task, Amber’s discussion of outliers is mostly correct, but the language she uses is again unclear. She writes, “The boxplot for Group 1 has four outliers, 34 and 54 are between 95-100 months, 8 are between 100-110 months, and 23 are between 110-120 months. The boxplot for Group 2 has one outlier of 178 between 100-110 months.” In this task, she correctly refers to the outliers as such, rather than observations, and the units are correctly identified as months. However, the way she lists the observation numbers, particularly 8 and 23, makes it sound like she is claiming they refer to the number of outliers in the interval. This is supported in the interview when Amber is asked about the boxplot and she says they are the number of outliers. Her confusion about what the numbers on the boxplot represent is a setback from the third task, but it is still an improvement from the second task.

In summary, Amber experiences mixed success in her characterization of distribution. There is no change in the way she discusses distribution shape, since in each task she identifies the shape without elaboration. She does make some progress with outliers, but struggles to read them from the boxplot and discuss them with clarity in her argument. By the end of the teaching experiment, her ability to identify and discuss outliers have improved, but are still at the emergent phase.

5.1.2.3 HYPOTHESIS TESTING

In this section, Amber's use of hypothesis tests, including evaluation of assumptions and calculation and interpretation of effect size, will be discussed. In this teaching experiment, there is only one opportunity to examine scaffolding of hypothesis testing, and it is between the third and fourth tasks.

In the third task, Amber's evaluation of assumptions for the t-test consists of simply stating that they have been met. She does not support the statement with a mention of the sample size or the distribution shape. It is given in the last sentence of the argument as if it is a part of the conclusion, whereas it is usually presented prior to the hypothesis test since it is a condition that must be met for the hypothesis test to be valid. Amber correctly states the hypotheses, p-value, and the conclusion of the test. Her interpretation of the conclusion is, "There is not sufficient evidence to conclude that the amount of claims the company must pay to repair or replace the refrigerators is less than $175 on average, and they will make a profit." It is unclear from her sentence of interpretation whether she is asserting the company will make a profit, or if the clause "they will make a profit" is included as part of what there is not sufficient evidence to demonstrate. However, in her interview, she clarifies that she means there is not sufficient evidence to show the company makes a profit. She states the effect size and identifies it as small. To interpret the effect size, she explains Cohen's guidelines for small, moderate, and large effect sizes, though she presents them much more rigidly than they were discussed in class. In the interview, Amber requires little guidance in discussing the hypothesis test. She quickly and correctly sets up the hypotheses and states the results of the test, easily answering the question of whether the test results in being unable to conclude the company makes a profit or being able to conclude the company does not make a profit. She has no trouble conducting a hypothesis test and she is able to incorporate it into a statistical argument, albeit using basic language.

In the fourth task, there are minor changes in Amber's use of hypothesis testing. She evaluates assumptions for the hypothesis test near the beginning of the argument. The placement of the evaluation of assumptions makes more sense in this task than in the previous one; it is part of the statement of sample sizes with the other descriptive statistics, and it is prior to the hypothesis test. Amber again correctly states the hypotheses, p-value, and conclusion for the t-test. Her interpretation for the hypothesis test conclusion is, "There is statistically significant enough evidence to conclude that Caucasian offenders will get more lenient sentences than any other group." The conclusion is correct and does not include the confusing clause at the end of the sentence as in the third task. However, the statistical language is not consistent with the language that was used in class, where results were described as either statistically significant or not statistically significant. Amber next gives the effect size, listing Cohen's guidelines using the same language as she did in the third task. Finally, she interprets the effect size as there being little difference "between the data"; she is correct that the effect size is small, but the language of referring to the groups simply as "data" lacks clarity. In the interview, Amber has lost some of the fluency in discussing hypothesis testing from the previous task; she confuses the null and alternative hypotheses and does not immediately state the correct conclusion. Her progress from the third task is thus mixed; she appears to have a better understanding of evaluating assumptions, yet she loses some of the ease of discussing hypothesis testing.

Overall, Amber's grasp of hypothesis testing seems to be stronger than her understanding of some other descriptive statistics, though she experiences a loss of fluency in discussing it between the third and fourth tasks. Like the other elements, even though she appears to understand hypothesis testing, she struggles to incorporate it in a clear way.

5.1.2.4 CONTEXT IN ARTICULATING RESULTS

In this section, context in articulation of results is examined as it relates to center, spread, distribution, and hypothesis tests. Those elements are not present in the first task, so this section focuses on the second, third, and fourth tasks.

In the second task, Amber uses the colors of the cars (red, blue, and neutral) in stating the descriptive statistics, a link to context. She attempts to use the units of the problem, but her usage of the word "units" instead of miles per hour makes it difficult to consider the results in context. She does not go further to use the results to make meaning of the scenario, not even to answer which color car has the fastest speed.

In the third task, Amber’s use of context in articulation of results is still incomplete. She uses the correct units, dollars, in reporting the mean and standard deviation, an improvement from the second task. However, as was previously mentioned, when discussing outliers, she refers to them as “claims paid.” That, and the accompanying interview (see section 5.1.1) about what each observation represents reveal an incomplete understanding about the intended context. Still, in the hypothesis test, she successfully states the conclusion in context, saying “There is not sufficient evidence to conclude that the amount of claims the company must pay to repair or replace the refrigerators is less than $175 on average, and they will make a profit.” Here the requirement in the task to interpret the conclusion of the hypothesis test facilitates placing the results in context. She does not attempt to link the effect size to context, interpreting it only as a “small effect.”

In the fourth task, Amber more consistently uses context in articulation of results. She uses correct units, months, to relate the means, standard deviations, and outliers. Correct usage of units in discussing outliers is an improvement from the third task. Again, in the hypothesis test, she states the conclusion in context, “There is statistically significant enough evidence to conclude that Caucasian offenders will get more lenient sentences than any other group.” Unlike the third task, here she does attempt to link effect size to context, saying “A small effect size means that there is hardly any difference between the data we have been given for this research.” It is a weak link to context since she does not specifically state the groups or the implications of a small effect size in this scenario; still, it is an improvement from the third task.

Over the course of the teaching experiment, Amber incorporates context in increasingly meaningful ways into the articulation of results. Hindered in the second and third tasks by confusion about the scenarios, by the final task she relates the results to context, using units correctly. In the final two tasks, statements of interpretation for the hypothesis tests aid in situating the results in context.

5.1.2.5 CONCLUSION

Throughout the teaching experiment, Amber shows growth in articulation of results in some, but not all, areas. She does not state the results in the first task, but in the subsequent three tasks she gives them ample attention. Considering individual elements of articulation, her discussion of center changes little; she states the measures of center comfortably in both the written arguments and interviews but does not compare them unless prompted. Her discussion of spread changes little in each of the written tasks, but in the interviews she goes from being uncertain even what measures or graphs to consider, to at the end of the teaching experiment being confident in choosing a measure. For distribution, the identification of distribution shape and outliers is mostly correct but unclear in the statistical arguments throughout the teaching experiment. By the end of the teaching experiment she is still not able to discuss them in the interviews. For hypothesis testing, between the third and fourth tasks her discussion of assumptions is moved to a more logical place in the argument, and her interpretations are linked more directly to context. The ease with which Amber discusses hypothesis testing declines, however, between the two tasks that include it. Overall, throughout the tasks, Amber articulates results correctly but without clarity in her writing.

5.1.3 MAKING INFERENCES

The final criteria of statistical arguments, making inferences, occurs primarily in two ways in each task. The first is identifying the presence of any sampling bias. The second is answering the research question, taking into account any sampling bias noted earlier in the argument. These elements are considered separately in this section.

5.1.3.1 EVALUATION OF SAMPLING PROCEDURES

In the first task, Amber identifies two problems with the sample. The first problem she observes is that “50 teenagers does not represent teenagers across the nation.” This shows she is considering the population, which she identifies as all teenagers in the United States. The second problem she describes is the presence of sampling bias; she determines the location in which data was collected not to be representative because it was all done in the same place of business in one city. She recommends sampling teens from varied public locations, and in several states and cities to introduce diversity in the sampling because it will “keep a random sample of teens across the world”. Though in this sentence she incorrectly uses the term *random sample,* it is another direct reference to the population; here she asserts the population is all teenagers worldwide, which contradicts her earlier assumption of the population as all teenagers nationwide. Still, her argument appears to be grounded in an understanding of the importance of having a representative sample.

In her second task, Amber couches the problems with sampling in recommendations for improvement. The first problem, sampling bias, is found in her recommendation for the sample to include data from more than one town. This is an implicit identification of sampling bias, since a more diverse sample is needed only if the given sample is not representative of the population. The second problem, a small sample, is found in her description of an advantage of her recommendation, “the study would be more beneficial because the study would be various between each town and possibly include more vehicles.” This statement implies the sample size is too small to be able to represent the population. These are the same two problems Amber identifies in the first task, but phrasing them as recommendations is more subtle language.

In the third task, Amber does not note any problems with sampling or data collection procedures. The intended context includes as a source of bias in the scenario that only the first customers who purchase the warranty are selected for the sample. However, this is a less obvious source of bias than the other scenarios, so Amber’s failure to notice it is not an indication of weakening ability in identifying bias.

In the fourth task, just as in the first two tasks, Amber again finds two flaws with sampling and data collection procedures, and they are the same two flaws. The first problem she identifies is a biased sampling method. She explicitly states the sample is not representative because it consists of only one type of felony. She suggests sampling all felonies, stratifying by race. Her recommendation, “The sampling and data collection procedure should consist of all convicts that committed a felony, and then sample that data by race,” implies she considers the population to be all individuals convicted of a felony. If this is the case, she has missed another source of bias in the study, the limitation of the sample to one state in the U.S. In the interview, she indicates she thinks sampling from one state is representative of the population, but she did not elaborate. The second flaw Amber finds is the sample size. She says a benefit of her recommended sampling method is that the sample size would be larger, implying she finds the samples in this scenario to be on the small side.

From examining Amber’s evaluation of sampling procedures, there is no notable change over time. Throughout the teaching experiment, she identifies some, but not all, sources of bias in sampling. She also finds most of the sample sizes to be too small. She presents her concerns about sampling and data collection procedures with varying specificity, at times stating them as flaws and at other times presenting them only as implications of her recommendations for improvement.

5.1.3.2 ANSWERS TO THE RESEARCH QUESTIONS

In her first task, Amber does not attempt to infer to the population. However, in the interview, when asked how she would answer the research question, given options such as “yes”, “no”, “yes, but…”, “no, but…” “cannot be determined” she says she would answer the research question with yes, but the survey should have asked about a topic relevant to teens. When asked if the flaws in the design render the data and resulting conclusions useless, she says no, the concerns with sampling and data collection don’t warrant throwing out the entire dataset. While she does not answer the research question in her written argument, it appears that when prompted, Amber is able to answer the research question for the overall population.

In the second task, Amber’s summary statement at the end of the written task does not address the research question. She reiterates that there are problems with sampling and data collection, and then simply concludes that enough information is provided to answer the question. Since there is no summary in the first task, this is an improvement, but it falls short of making a true inference to the population. The following exchange from the interview gives more insight into her understanding of inference:

> TR: OK, so what then would you conclude about the population from all this?
> Amber: It doesn't really matter on the color of the vehicle because all of them were speeding it's not just what color the car is, it depends on the driver because all the groups are kind of close to each other. It's not that far off.
> TR: And what population would you infer to?
> Amber: Like which?
> TR: Are we talking people? People in the US?
> Amber: I would say all people in that town, like if they were to do Charlotte, only people in Charlotte because you can't say around the world because they only did it in one town it's not an accurate data.
> TR: And to people in all speed zones or only certain ones?
> Amber: I think only in certain speed zones.
> TR: Which ones?
> Amber: Ones that are 60 mph or ones that are like 75 mph—higher ones compared to lower ones. Because on the highway you have people at speed compared to what it's like 30 mph nobody's really going to do 60 in a 30—well, I would hope not.

In this exchange, she answers the research question and, with prompting, discusses limitations on the inference based on the sampling bias; i.e., the conclusions apply only to the geographic area and speed zone represented in the sample. Though Amber's written argument does not include inference, she is comfortable inferring to the population in the interview.

In the third task, for the first time in this teaching experiment, Amber infers to the population in her written argument. Her conclusion "the study did not provide sufficient amount of evidence to conclude that the company will make a profit if the amount of claims the company must pay to repair or replace the refrigerators is less than $175" is only a restatement of the result of the hypothesis test, but it is included as part of a summary paragraph designed to answer the research question. She accompanies the conclusion with a restatement of her recommendation for the scenario to focus only on one brand of refrigerator, but she does not integrate the recommendation into the inference.

In the fourth task, Amber's inference is more thorough. She ends her argument with a summary paragraph in which she addresses the research question. She reiterates that she has concerns with the sampling procedure, yet as the following passage shows, she concludes the answer to the research question is yes, Caucasian offenders receive more lenient sentences.

> Even though I feel like the research sample and data collection procedure could change, the research still included enough evidence to conclude that Caucasian offenders will get more lenient sentencing than any other group.

Unlike the third task, this inference does synthesize, however subtly, Amber's reservations about sampling. Her use of "even though" and "still" reflect that she is considering these concerns, but does not view them as sufficiently problematic to interfere with the conclusion. Ideally, she would be more specific about these concerns and the resulting limitations on the conclusions, so there is still room for improvement.

5.1.4 CONCLUSION

Amber's work was chosen for analysis in part because she reports that the course and these assignments were difficult for her. Her arguments show she struggles to understand context, which makes it difficult for her to identify sampling bias and answer the research questions. Spread and outliers present particular challenges for her. However, she shows improvement in many of these areas over the course of the semester. It is impossible to tell the extent to which her growth is due to the scaffolding of the tasks or her improved health, but she reports being pleased with her progress. At the end of the teaching experiment, she says these tasks have not only helped reinforce the course concepts, but they have provided examples of how statistical analysis may be used in the real world.

5.2 CASE STUDY 2: LEAH

The subject of the second case study, Leah, is an adult student in her mid-40s who is Caucasian and from the Southeastern United States. She is taking classes part-time and has accumulated credits equivalent to a second-semester sophomore. She is returning to school to study nursing, after being a stay-at-home parent for several years. Before the birth of her children, Leah worked as a paralegal. From discussions with Leah, it seems she has a good grasp on course concepts, but her test grades do not reflect her understanding because she struggles with calculations. Her work was chosen for analysis because it shows how statistical argumentation can reveal understanding not captured by calculation-based problem solving. Her work was also chosen because Leah has direct experience with some of the scenarios provided to students in this study, so her arguments yield insight into the effects of familiar contexts on statistical argumentation.

5.2.1 LINKING TO CONTEXT

In the first task, one way Leah links to context is in evaluating the appropriateness of the survey question, which is “Would you support suffrage for 16- and 17-year olds?” to address the research question, which is whether teens pay attention to local and national politics. She finds them to be not in alignment and points to the difference between paying attention to politics and the desire to vote. Her observation about the lack of alignment between the survey question and research question shows her context to be consistent with the intended context. She then proposes a survey question more directly relevant to the research question. She explains that her recommended survey question "Do you read or watch local and national news weekly?" would answer the research question because local and national news includes coverage of political issues. Her proposed survey question and her justification of it both rely on her knowledge of the context.

A second way Leah links to context is given as part of her evaluation of sampling and data collection methods. She writes,

> For one, I don't think many 16 or 17 year olds know what the term suffrage means. The organizer assumes the teenager is familiar with political terms. This would definitely skew the participant's response to not being accurate or not responding at all, because they don't understand the question that is being asked. The data that was collected makes me believe that this was true. Seventy six percent of the participants responded no to their right to vote. This is not reasonable to me. The teenagers I know want every opportunity to be treated as an adult. I would think more participants would have voted yes to supporting suffrage.

This passage shows a three-step thought process in which Leah interacts with context. First, she uses her previous experience with teenagers to suspect respondents misunderstood the survey question, in particular the political term *suffrage*. Second, she conjectures how the results might be impacted as a result of respondents misunderstanding the question. Though she does not state her conjecture, it is implied by her next observation, the third step in her thought process. Third, she compares the results given in the scenario to her conjecture and her knowledge of what respondents' opinions are likely to be, and she finds the results make more sense if respondents misunderstood the question. This is a deep interaction with context, and it shows her context is consistent with the intended context, in which the word *suffrage* sounds negative and causes respondents to say they are opposed to it.

In the second task, Leah again uses context to evaluate sampling and data collection procedures, saying the sample may be biased because the data is collected by police officers who, according to the hypothesis described with the research question, pay more attention to red and black cars. For this reason, she recommends the data be collected by a person not connected to law enforcement. This is consistent with the intended context, in which the presence of law enforcement may cause people to drive more slowly. Later in the argument, Leah uses the context of the study to explain the importance of having a geographically diverse sample; she says traffic patterns and tolerances of car speeds vary from place to place, so data should be collected from more than one city. This is again consistent with the intended context. Leah interacts with the context less in the second task than in the first task, but she still uses the context effectively to support her conclusions and recommendations about sampling and data collection procedures.

In the third task, Leah delves even more deeply into the context than was intended. She has direct experience with this scenario; in the interview, she talks about a time she purchased a refrigerator and considered the cost of repairs as a factor in choosing which one to buy. One way she uses context is in evaluating the sampling method. Leah identifies it as a convenience sample, which is often undesirable for a study, but she concludes, “This is not particularly bad because the company is wanting to use the data for their own benefit (we’re assuming), so trying to get data from another location or some other way would not be beneficial.” By considering the sampling method in context, Leah is able to determine that in this case a convenience sample can be a useful way to collect data.

A second way Leah uses context in the third task is by evaluating the research question itself. The research question is “Does an appliance company make a profit when they offer a two-year warranty on a certain brand of refrigerator?” In the intended context, refrigerators of the same brand are similar in quality. Leah recommends limiting the sample to one model of the brand because in her context, the model impacts quality and the need for repairs. She contends higher end models have fewer problems but more expensive parts than lower end models, both of which impact repair costs. Leah goes further to say that her recommendation of analyzing the models separately could help the company determine which models are profitable. While this is a possible benefit of her recommendation, the task called for students to address the given research question rather than proposing a new one. Still, it shows a deep interaction with the context of the study.

A third way Leah links to context in the third task is in her recommendation that in order to achieve credibility the study needs to be conducted by a third party company. To support her recommendation, she describes a scenario in which a disgruntled employee interferes with the data in an attempt to undermine the company’s profitability. In her interview for this task, Leah says she saw a similar scenario occur at a place she once worked. This is an example of a student’s personal history directly impacting her context of the study, which is more likely to happen with adult students who have accumulated more life experiences than traditionally aged students. While more familiarity with the context of a study may enhance an argument, in this argument it verges on becoming a distraction. Leah has described a hypothetical scenario which according to her is far-fetched, and used it to recommend hiring consultants to conduct the study on behalf of the company, at a potentially high cost. As part of the same paragraph, she argues it is important to get accurate data “to ensure profitability.” The research question in this scenario is whether the company makes a profit on the warranty; it is not specifically to create profit for the company. While this may be a subtle point, it is raised again in the interview during a discussion of outliers, which were all on the upper end of the data. When asked if the outliers should be included in the dataset, Leah replies “no,” they affect profitability and if she is a manager, she does not want them in the dataset. When pressed about whether it is important to use them to determine whether the company is making a profit on the warranty, she replies the outliers “should not occur in the first place” because they cost the company money. Leah seems to lose sight of what the research question is, focusing instead on how to improve profitability for the company. In this instance, Leah has become so focused on the context that she is no longer making the same argument. Leah is clearly considering context throughout the third task, even to the extent that it interferes with the effectiveness of the argument.

In the fourth task, Leah again focuses heavily on the context of the study. As a former paralegal, she has direct experience with this scenario. She first expresses concern about the source of the study. While in the intended context, a criminologist is an academic researcher who ideally has little bias, in Leah’s context this is not necessarily the case. In particular, according to Leah’s context, the race of the researcher is a factor that may cause bias. Her recommendation, as it was in the second and third tasks, is to hire a company to conduct the study. In her context, though a company is comprised of individuals of varying races, it is detached enough to generate unbiased results. In this task, while Leah links to context frequently, she does not discuss it so deeply that she veers from her overall argument.

Throughout the teaching experiment, Leah incorporates context throughout her arguments. It is not possible to say her use of context improves, because she effectively incorporates it into the first task, leaving little room for improvement. In one instance, Leah’s experience with the context becomes a hindrance, as she seems to get so focused on it that she loses track of the argument. Still, overall, Leah’s arguments show how an adult student’s experience with context can enhance statistical arguments.

5.2.2 ARTICULATING RESULTS

In this section, the way Leah articulates results in her statistical arguments will be examined, using her work and interviews from the second, third, and fourth tasks. The first task includes only one result to report, a percentage, so it is not included.

5.2.2.1 CENTER AND SPREAD

In the second task, Leah uses the skewed nature of the distributions to choose the median as the best measure of center. This is consistent with what students were shown in class, that for skewed distributions, the median is generally preferable to the mean as a measure of center since the median is less influenced by outliers. The use of distribution to choose a measure of center is a combination of elements, a desirable trait of statistical argumentation. Leah states the values of all three group medians and compares them, which is important for addressing the research question. Leah continues by discussing the spread of each group. She simultaneously states and compares the standard deviations, finding the neutral cars to have the largest spread, followed by the black cars and finally the red cars. She then explains why the result seems to her to be counterintuitive. She writes,

> This is unusual to me because normally the bigger the sample size the less variability. In this case, the neutral cars are the largest sample and have the highest variability. I feel that the samples should have been larger and all the same size to have less variability.

The misconception that large sample sizes cause smaller standard deviations is a common one shared with several other students in this task. They are apparently confusing margin of error, which does decrease for larger sample sizes, with standard deviation, which does not. Indeed, by the time students completed this task, confidence intervals were being covered in class. The decision was made not to address this fallacy in the interview to allow for focus on other aspects of the argument; instead, the concept was discussed with the class as a whole. Other than the misconception about standard deviation, Leah's discussion of center and spread in the second task is correct and thorough; in addition to the aforementioned use of distribution shape to choose a measure of center, she states and compares measures of center and spread.

In the third task, Leah does not get into as much depth with center and spread. Aside from the discussion of the t-test later in the argument, her statement of center and spread consists only of a sentence stating the mean and standard deviation. She does not explain the choice of mean as measure of center, even though the distribution of this data is more skewed than the distributions in the second task. Not using the distribution to choose a measure of center is the only substantive change in Leah's discussion of center and spread between the second and third tasks, though, since most of her discussion of center and spread in the second task consist of comparing across groups, and this data does not contain groupings.

In the fourth task, Leah begins discussing center and spread by stating the means for both groups. She compares the means and evaluates the magnitude of the comparison as “not a large center difference.” Next, she lists the standard deviations and interprets them as showing “group 1 having a greater distance from the mean than group 2.” This sentence indicates Leah understands the concept of standard deviation as a measurement of the distance from the mean of observations in the dataset. In this task, Leah does not use the distribution shapes to choose which measure of center to report. However, near the end of her argument as she is summarizing her findings (the statistically significant difference in means with small effect size) she points out that there are four outliers in the Caucasian group sentences and they “could have affected the results.” Though it is stated vaguely, this is evidence that she is considering the relationship between distribution and center.

Leah’s discussion of center and spread in articulation of results reveal that, for the most part, she does appear to understand the two concepts and use them effectively in her arguments. Throughout the second, third, and fourth tasks, she chooses appropriate measures for center and spread, and then states and interprets them correctly. By the fourth task, she describes spread conceptually as distance from the mean. She provides less detail in the third task than in the second task or the fourth task, which is easily explained by the lack of grouping variable. In the second and fourth tasks, she shows evidence of considering the relationship between center and distribution.

5.2.2.2 DISTRIBUTION

In the second task, Leah addresses distribution by labeling the distribution shapes and listing outliers. She characterizes black cars as having a normal distribution, and neutral and red cars as having skewed distributions; this statement is included as part of her sentence explaining why she chose the median as the measure of spread. She does not specify which direction the skewness is. None of the three histograms closely resemble the common histogram shapes (normal, negatively skewed, and positively skewed), so any attempt to categorize them is somewhat subjective. Still, the histograms for red and black cars are very similar in shape, so it is curious Leah would label them differently. For outliers, she lists the approximate value of each outlier, which group it is in, and whether it is on the high or low side of the data. Her characterization of distribution is correct and written in a way that is easy to understand.

In the third task, Leah provides substantially more detail about the distribution. She labels the distribution shape as right-skewed; in this task, the shape closely fits a positively skewed distribution, so her characterization is correct. In this task, she identifies the shape of the distribution in a sentence evaluating assumptions for the t-test. For outliers, she lists information that is only available by cross-referencing the histogram and boxplot. Because there are so many outliers (a total of 18), the observation numbers are not all listed on the boxplot. Additionally, some of the values of the outliers are repeated, making it impossible from looking at the boxplot alone to determine how many observations are represented by each value. Leah approximates each value by examining the boxplot. Next, she determines, based on the histogram, the number of outliers there are at each value. She estimates there to be 15 outliers of $850, two outliers of $600, and two outliers of $550. Her estimates are very close; fourteen outliers are between $800 and $850, two are between $600 and $650, and two are between $500 and $550. Using the two graphs together to determine not only the values, but also the frequencies of the outliers, shows a deep understanding of both boxplots and histograms. There are fewer outliers in the other tasks and they are able to be counted from the boxplots, so it is not possible to determine whether Leah cross-references the boxplots and histograms before or after this task.

In the fourth task, Leah provides different details about the distribution. She correctly identifies the distributions for both Caucasian and non-Caucasian offenders as being positively skewed. She states the minimum and maximum data points for each group. This is the first task in which she lists the extrema. For outliers, she lists the number but not the values of outliers in each group, which is less information than in previous arguments. Stating the extrema is more detail, but at the same time, she provides less information about the outliers. Overall, her observations are still correct and meet the requirements of the task.

Considering Leah’s discussion of distribution over the course of the teaching experiment, she provides varying levels of detail regarding distribution shape and outliers. Her combined interpretation of the histogram and boxplot in the third task is noteworthy. No clear trend emerges in her progress; instead, she appears to have a solid understanding and ability to incorporate distribution into her arguments throughout the teaching experiment.

5.2.2.3 HYPOTHESIS TESTING

In the third task, toward conducting the hypothesis test, Leah first evaluates assumptions. She indicates the t-test can be used because the sample size is large enough, despite the skewness shown in the histogram. Her statement is correct, but it is included as part of stating the sample size and distribution of the data. Usually test assumptions are discussed immediately before the hypothesis test, so this sentence is a bit out of place in her argument. Later in the argument, Leah conducts the hypothesis test. She states the hypotheses in symbols accompanied by units. After stating the p-value and comparing it to the level of significance, she correctly states and interprets the conclusion of the test. She then finds the effect size, including a full description of the steps in the calculation. Consistent with Cohen's (1988) guidelines, she evaluates the effect size as small and the result as not being practically significant. She does not otherwise connect the effect size to the result of the t-test. Her conclusion, "there is insufficient evidence that μ<175 dollars claimed" is stated correctly. However, later in the argument, she changes the phrasing to "We found that the data proved that the company did not spend on average less than 175.00 dollars on a claim." Instead of the actual result of failing to reject the null hypothesis, she mistakenly claims that the null hypothesis is correct. She states it correctly in the interview, so this appears to be a momentary lapse. Overall, with the exception of overstating the conclusion at the end of the argument, the elements associated with the hypothesis test are all correct and appropriately used in the argument.

In the fourth task, Leah approaches the hypothesis test first by evaluating assumptions for the t-test, finding that they have been met via sufficiently large sample sizes. In the next sentence, she writes the hypotheses using symbols μ_1 and μ_2, then describes in detail how she found the p-value, and states the rejection rule for the test. Labeling the groups as 1 and 2 in this scenario is correct, but it may have been more effective if she had labeled them in context as Caucasian and non-Caucasian. She states the conclusion of the hypothesis test, describes the result as statistically significant, and interprets it in the context of the problem. Moving next to effect size, Leah describes in detail the steps for calculating for Cohen's *d*. She states the value of Cohen's *d* and interprets its magnitude as small. She then combines the result of the hypothesis test with the effect size to conclude that it shows the difference between statistical significance and practical significance.

Between the third and fourth tasks, Leah's incorporation of the hypothesis test improves in three ways. First, the assumptions are stated immediately before the hypothesis test. Second, the interpretation is stated correctly throughout the argument. Third, the result of the hypothesis test is interpreted in conjunction with the effect size, providing a limitation to the conclusion of the hypothesis test.

5.2.2.4 CONTEXT IN ARTICULATING RESULTS

In the second task, Leah incorporates context throughout her articulation of results. As she states the descriptive statistics, she labels them by color and uses miles per hour as units. She identifies which groups have the highest and lowest medians, and which group has the highest variability. Specifying that she is attempting to explain some of the variability, she postulates that the cars may have been clocked in different areas, some of which drivers are more likely to exceed the speed limit. She also attempts to explain the outlier for neutral cars (but not the outlier for black cars) in context by speculating that the car may have been clocked in a different area than the other neutral cars. In this task she links every element of articulation of results—center, spread, and distribution—to context, first by describing it and then by explaining what it means in context.

In the third task, Leah again uses appropriate units, dollars, as she discusses the center and spread of the distribution. However, she also includes the sentence, “This study does use appropriate units to describe the data, using claims paid (dollars).” She appears to have interpreted the “uses units where appropriate” criteria on the list of elements to include in statistical arguments as requiring an evaluation of units; this is a fairly minor misinterpretation and common to several other students, despite repeated clarifications of the element. As she is discussing distribution, she uses the high number of outliers to conclude they are not the result of measurement error. When she discusses the hypothesis test, her initial conclusion “I conclude that there is insufficient evidence that u<175 dollars claimed” and observation of the effect size as “small” and “not practically significant” are less clearly linked to context than other aspects of her argument. However, she later explains the results in the words of the problem, even using the results to recommend the company increase the price of the warranty. Overall, on the third task, though she is apparently confused about the “units” criteria, Leah frequently includes considerations of the practical implications of the data analysis on context.

In the fourth task, Leah again evaluates the units as being appropriate for the study, showing she apparently still has the same misconception about the “uses units where appropriate” criteria as she has in the third task. Still, in this task she uses the units—months in this case—when stating descriptive statistics, though when she refers to the groups, she calls them groups 1 and 2 rather than Caucasian and non-Caucasian. Though she does not attempt to explain outliers as she did in the second and third tasks, she does explain some of the variability by listing factors besides an offender's race that may impact sentence length. In the hypothesis test, Leah does not state the hypotheses in context but she does state the result of the hypothesis test and the effect size in context, saying non-Caucasian offenders have a higher mean sentence length than Caucasian offenders but that the difference is not enough for the result to be practically significant. In the fourth task, though some of the description is less directly linked to context, it is still incorporated into every aspect of articulation of results.

Overall, Leah's arguments throughout the teaching experiment are thoroughly grounded in context. She attempts to relate nearly every result to context and compares it to her real-world experiences. There does not appear to be growth, but her initial work is of high enough quality that there is not much room for improvement.

5.2.3 MAKING INFERENCES

The final criteria of statistical arguments, making inferences, is examined in this section. As in the first case study, this section is divided into two parts: evaluation of sampling procedures and answers to the research question.

5.2.3.1 EVALUATION OF SAMPLING PROCEDURES

In the first task, Leah puts considerable emphasis in her argument on the definition of the population and the lack of representative sample in the given scenario. Because the population is not stated in the information provided, she believes it may be all teens worldwide, which she assesses as too broad and recommends limiting to the United States. She evaluates the sample described in the scenario to be biased in three ways: the survey is only conducted in one location, the sample size is too small, and the respondents are not randomly chosen. Leah specifies these problems make the sample not represent the population, and suggests instead using simple random sampling to collect data, which she claims “enhances the accuracy of the data.” The words *accuracy* and *validity* were not covered in class the way they might have been in a research methods class, so Leah is not expected to know the difference between them; in this case, her use of the word *accuracy* appears to refer to the representativeness of the sample. Toward a more representative sample, she recommends randomly sampling one large high school for each region in the United States and then randomly selecting at least 250 16- and 17-year olds to “strengthen the representation of the population of all teens in the U.S.” Though her recommended sampling procedure is biased toward students attending large high schools and is limited to 16- and 17-year olds, it is a substantial improvement over the one described in the scenario. In this task, Leah clearly demonstrates understanding that the goal of the study is to infer whether the sample is representative of the population; she identifies the population, evaluates the given sampling procedures, and when she finds them to be biased, recommends new procedures more likely to yield a representative sample.

In the second task, Leah again emphasizes in her argument the importance of a representative sample. She identifies three main problems with the procedure for sampling and data collection, and for each problem she recommends a remedy. First, she says bias may be introduced by a police officer being the person collecting data, and points to the statement included in the second task as evidence, “Some people believe that drivers of red and black cars speed more often than other drivers, causing police to watch those drivers more closely.” To remedy the bias, she suggests the person collecting data should have no connection to law enforcement. Second, she judges the sample sizes to be too small, and recommends every color of car contain at least 30 observations. Third, she finds bias in the sampling being limited to one city because drivers in other cities may not behave the same way. To remedy the bias, she recommends random sampling cities across the U.S., making sure to include data from each region of the country, which implies the population is all cities in the United States. In the second task Leah refers repeatedly to the population and whether the sample is representative of it. She appears to be considering inference consistently.

In the third task, Leah finds fewer problems with sampling and data collection. She identifies the sample as being one of convenience, but says a convenience sample is acceptable in this case since it is only to be used for the company’s internal purposes and it would not be useful for them to collect data from other companies. The primary problem she finds is the risk associated with having an employee of the company collect the data, since the employee may want to undermine the company’s profitability. To reduce this type of risk, Leah recommends an “outside reputable professional survey company” to collect the data. The issues with sampling and data collection in this task are minor; Leah nonetheless identifies the convenience sample and evaluates it as being not problematic under the scenario.

In the fourth task, Leah discusses two concerns about the study. Her first concern is uncertainty about the researcher; she writes that the race and employer of the researcher are factors that may cause bias. To remedy the possible bias, she suggests an outside firm be contracted to collect the data. Her second concern is the narrow sampling frame. She recommends the sample consist of more than one state and more than one felony. She says the population is all felons who have been sentenced, and though she does not directly address geographic location, her recommendation of using "randomly picked states" implies she is thinking about the U.S. She also recommends using various felonies to better represent the population. Though Leah's assertion that an outside company would be unbiased may be easily disputed, in the fourth task, she still considers the population throughout the argument, identifies sampling bias, and proposes solutions to reduce sampling bias.

In conclusion, Leah effectively identifies sampling bias throughout the teaching experiment. She is quick to suspect bias on the part of the researchers, but in every case she justifies her concerns. In every task, she shows she understands the importance of having a sample that represents the population, which is vital to making statistical inferences. There is not clear indication of change over time; instead, similar to other elements of Leah's arguments, they are strong in the first task and remain so until the fourth task.

5.2.3.2 ANSWERS TO THE RESEARCH QUESTIONS

In the first task, Leah's argument focuses on evaluating the information given in the scenario, and she does not attempt to answer the research question. In the interview, when asked what the answer to the research question would be, she answers yes, teens do pay attention to local and national politics. When pressed whether her answer is based on the survey results or her own experience, she says it is based on her general knowledge rather than the survey results, because the survey is too flawed to get an answer from it. Though she does not attempt to answer the research question in the written task, Leah is able to infer to the population in the interview and justify her decision not to attempt to answer it in the written task.

On the second task, Leah again does not attempt to answer the research question. She compares the sample medians in the descriptive statistics section of the argument, with the statement “Neutral colored cars have the lowest median mph with 67.00 and red colored cars have the highest median mph with 76.00,” but she does not directly answer the research question or infer to the population. In the interview, she answers the research question in the affirmative, saying there is a difference in the average speeds of drivers of the three car colors, but she tempers her conclusion by saying the data collection is flawed so the answer is not compelling to the average person. She has no trouble answering the research question in the interview, even though she does not think the answer is strongly supported by the data collected.

In the third task, for the first time Leah attempts to answer the research question in writing. She addresses the research question after discussing the hypothesis test. She writes,

> We found that the data proved that the company did not spend on average less than 175.00 dollars on a claim. The effect of the warranty in regards to profitability for the company is very small. We can infer that the warranty price needs to be higher or that the data collection procedures need to be more detailed to recognize where profitability, if any, lies.

Her answer is based on the hypothesis test and effect size, which is not surprising since these are inferential statistical methods. Though she overstates the conclusion from the hypothesis test, it is still an answer to the research question. Her answer includes not only the negative answer; it reiterates the effect size—even though no effect was actually found—as small, and it goes further to make recommendations for the company. The effect size is not integrated into the conclusion of the hypothesis test, meaning it is not stated in a way that either limits or extends the conclusion; rather, it is stated separately in addition to the conclusion. In Leah's summary paragraph at the end of the argument, instead of discussing the answer to the research question, she observes that the data fits the research question, then concludes the "data is unreliable" because of flaws in the data collection procedure. The reiteration of sampling bias is part of Leah's conclusion but it is stated separately from the answer to the research question. It is not synthesized as it would be if it were stated as a limitation on the conclusion.

In the fourth task, Leah again answers the research question with a focus on the hypothesis test and effect size. She integrates the effect size into the results from the hypothesis test, as shown in her statement, "Even though Non-Caucasians have a higher average mean for sentences received, those sentences were not high enough over Caucasians mean to practically support the study." This sentence shows she is able to determine that the Caucasian offenders get lower sentences on average than non-Caucasian offenders, but she uses the small effect size to mitigate the result of the hypothesis test. Later in the argument, she reiterates the flaws in the study, and while she does not directly use them to inform the answer to the research question, she does speculate that if the flaws were remedied, the effect size may be larger. This is an improvement from the third task, in which she does not directly connect the flaws of the study to the research question.

Over the course of the teaching experiment, Leah improves in answering research questions. In the first two tasks, she does not attempt to answer the research questions, focusing solely on evaluating sampling and data collection methods. In the third and fourth tasks, she answers the research questions, placing particular emphasis on the hypothesis test and effect size. She improves in the fourth task by synthesizing the results of the hypothesis test with the effect size and by connecting the sampling bias to the effect size, neither of which are done in the third task.

5.2.4 CONCLUSION

Leah's statistical arguments show an example of how an adult student may perform on the tasks in this teaching experiment. Her work is very strongly grounded in context. She attempts to link every aspect of the study—the sampling and data collection methods, the results, and the conclusion—to the scenario given. She examines sampling and data collection methods with a critical eye, finding bias and recommending remedies where needed. She frequently evaluates the results against her experience and common knowledge to see if they are credible, and she uses the conclusions to suggest implications for the scenarios. She improves over time in making inferences, as her arguments shift from being assessments of the quality of the scenarios to being true statistical arguments that address a research question given all the information provided.

5.3 CASE STUDY 3: KURT

The subject of the third case study, Kurt, is an international student. He is fluent in English though he is not a native speaker, having studied the language starting in middle school. A sophomore at the time of this study, Kurt had been in the U.S. for just over three semesters; thus, he has limited knowledge of U.S. culture. Kurt's performance in the course overall was superb, indicating he possesses both an ability to perform calculations and a strong understanding of concepts. His work was chosen for analysis in part because it provides an example of the statistical arguments a non-native English speaker can generate, a valuable perspective given the language intensive nature of these assignments. Additionally, of the participants in the study, his is the best work overall, so it shows an example of an excellent set of statistical arguments.

5.3.1 LINKING TO CONTEXT

Kurt begins the first task—and each of the four tasks in the teaching experiment—by summarizing the scenario given. Beginning a statistical argument with a restatement of the scenario is not presented in the list of elements to include in any of the arguments and it is not included in any of the example arguments. However, it is an effective way to begin the argument for two reasons: first, it presents the argument without the assumption that the reader has seen the scenario, and second, it begins the argument with a direct link to the context of the study. Thus, Kurt improves upon the example arguments presented to students in class.

Continuing in the first task, Kurt evaluates the sampling and data collection procedures provided in the scenario. He first notes a lack of alignment between the research question and survey question, listing three possible explanations of affirmative answers to the survey question that do not indicate the respondent pays attention to politics. In each of these alternate reasons for "yes" answers to the survey question—the respondent's desire to have the opportunity to vote, the leading nature of the question, and the respondent's desire to make a favorable impression on the authority figure asking the question—Kurt relies on his understanding of the scenario being discussed. Kurt’s observations indicate he believes young people view the right to vote as empowering and a privilege that causes them to feel like adults. He also believes young people perceive that adults want them to be interested in voting, since he says this is the answer that will impress the authority figure. Each of these aspects of Kurt’s context are consistent with the intended context.

In Kurt’s list of problems with the sampling and data collection procedures, he does not mention the wording of the survey question. The survey question described in the scenario includes the word *suffrage*. In the intended context, the word *suffrage* may be unfamiliar to some people; in particular, it may be confused with the word *suffering*. In Kurt’s interview after the first argument, when asked if he thought teenagers would know the meaning of the word *suffrage*, he responds that being new to the United States, he does not have a sense of how common it would be. Kurt’s context is directly shaped by his previous experience, much of which is outside the United States.

A final way in which Kurt uses context in the first task is in evaluating the credibility of the results. He writes,

> The result of the research study is theoretically reasonable. 24 percent of the sampled individuals said that they support suffrage for 16- and 17-year olds and 76 percent are against it. It is generally known that teenagers are not really interested in politics.

When Kurt compares the results of the study to his expectations, he finds them to be credible and supports his conclusion with information about how he sees the context.

In the second task, in addition to explaining the scenario, Kurt makes use of context in evaluating sampling and data collection methods. In particular, he uses it to justify the importance of the sample to contain a variety of cities and speed zones. He explains, “Drivers are probably more likely to speed in different speed zones compared to others.” In this aspect, his context is revealed to be similar to the intended context. In this task there is less reference to context, it is still present in the description of the scenario and the evaluation of data collection methods.

In the third task, after summarizing the scenario, Kurt uses context to evaluate the sampling and data collection procedures. He concludes the research question and data collection procedures “make sense” for the scenario. He subsequently interprets “make sense” to mean that the sampling and data collection procedures align with the research question and will yield an answer to it, which indicates he is considering the context. Kurt evaluates the sampling procedure favorably, explaining that there is no obvious demographic bias in the sampling procedure. He does not mention other kinds of bias that were part of the intended context, such as systematic differences in consumer behavior or quality of the refrigerators between the first products sold and later ones, but these were subtle aspects of the context. These two ways of using context—to summarize the scenario at the beginning of the argument and to evaluate the sampling and data collection procedures—are the same as in the second task.

In the fourth task, Kurt again uses context in the same two ways. After restating the information given, he uses context to discuss sampling bias. He justifies the need for the sample to include more than just one type of felony by explaining, “The degree of biased judgment could differ with the nature of the crime.” He also uses context to recommend dividing the non-Caucasian group into smaller groups, pointing out that the amount of bias could vary by race. Both of these observations are consistent with the intended context.

It may appear that Kurt places more emphasis on context in the first task than in the other tasks. However, the first task includes more problems with sampling and data collection than the other tasks; thus, it lends itself to more discussion of context. Additionally, much of the links to context in the second, third, and fourth tasks occur in articulation of results, which is discussed in more detail in section 5.3.2.3. If Kurt seems to place less emphasis on context in these tasks, it is at least in part because of the way the tasks are written and analyzed.

5.3.2 ARTICULATING RESULTS

In this section, Kurt's articulation of results are discussed. First, center, spread, and distribution are examined together, since in his arguments, Kurt combines the three concepts. Second, hypothesis testing is considered. Finally, his use of context in articulation of results is analyzed.

5.3.2.1 CENTER, SPREAD, AND DISTRIBUTION

In the second task, Kurt begins his discussion of center, spread, and distribution by using the lack of skewness in the histograms to justify the mean as the best measure of center. He states the means and compares them to each other, then compares the mean of the fastest group, the red cars, to the speed limit and observes the 16 mph difference as "remarkable." Kurt next addresses the spread, first by simultaneously stating and comparing the standard deviations of the groups. He uses the larger standard deviation to support his earlier recommendation that a larger sample size is needed. Kurt attempts to explain the higher standard deviation: "Neutral colored cars included a number of different colors which make the possibility for a spread higher. This underlines the statement that a higher sample size is needed for reliable results." He says the reason is for "reliable results," which he seems to interpret as having little variability. In combination with an assertion earlier in the argument that the sample sizes for each group should be the same, this statement reveals his conception that larger sample sizes yield smaller standard deviations. This conception was common enough among the students that it was addressed with the class as a whole rather than with students one-on-one.

As Kurt discusses distribution, he uses the boxplots to compare the ranges and medians at the same time, particularly noting the larger range of the neutral group. He identifies outliers for each group, listing the approximate values and whether they are on the high or low side of the data, but he does not attempt to evaluate whether they are legitimate data points. Kurt next discusses the implications of distribution on center and spread. He writes, "The fact that the graphs of the data of red and black colored cars is left-skewed means that there are more higher data points. So red and black cars are driven faster in average." He first cites left-skewed distributions for red and black cars as evidence of higher means than for the neutral cars. Later in the paragraph he uses an outlier in the black cars to explain why the black cars have a lower mean than the red cars. These two statements point to Kurt's use of the histograms and boxplots to interpret visually the mean as the balance point of the distribution. Also in this paragraph, in the sentence "The fact that the distribution of neutral colored cars is so high means that the data collected for those cars is not stating a specific trend," he seems to interpret the large range of data as lack of certainty about the mean. Since, by the time this assignment was submitted, confidence intervals were being covered in class, it is reasonable that students would confuse standard deviation with margin of error. In class lectures and discussion, connections among center, spread, and distribution, and multiple representations of each concept were often described, but for purposes of the argument tasks, students were asked only to describe distribution as shape and outliers. Kurt goes further, using distribution to confirm his previous conclusions about center and spread. This is noteworthy primarily because it is evidence that Kurt conceptualizes center and spread both visually on the graphs provided and mathematically as numerical summary measures. It is also evidence that he understands how center, spread, and distribution are interrelated. In terms of the argument, including these elements in the articulation of results helps explain the effect of some of the particular data points on the center and spread.

In the third task, Kurt begins articulation of results by describing the center. As in the second argument, he combines center and distribution, first by noting that the skewness in the distribution causes a large difference between mean and median, and then by using the skewness to choose the median as a measure of center. In his discussion of spread in the third task, Kurt again uses the word *reliability,* this time as a result of a large standard deviation. He writes,

> The standard deviation is fairly large, which is a negative factor for the reliability of the data collection. It basically means that there are probably a lot of outliers and that there are a lot of customers who have claims which are higher than the price they initially paid.

At first glance, the claim of poor reliability in the data collection makes it seem like Kurt believes sound methods should yield little spread. However, he follows it with an explanation, which indicates the variability among observations is a natural, not problematic, feature of the data. The use of the word reliability here as simply descriptive of a large spread in the data is different from his earlier use of the term as an assessment of the trustworthiness of the data. As Kurt discusses distribution, he again connects it to the reliability of the data, but after describing features of the distribution, he does not directly address how these features inform reliability. He lists the outliers, then uses them and the histogram shape to explain why the median was less than the mean for this data; this is another example of using multiple representations of center. Not only does Kurt compare the numeric measures of center, i.e., mean and median, he explains the difference using the histogram, a graphical representation of the data. His final observation about the distribution is to state what he calls the "regular range" of the data, which corresponds to the highest and lowest data points that are not outliers.

In the fourth task, as in the previous tasks, Kurt uses the distribution shape to choose a measure of center. In this case, skewness in the histogram causes him to use of the median, which he then states and compares, noting the magnitude of the difference between groups as small. Kurt next discusses spread by stating the standard deviations and comparing them to each other, concluding they are similar in magnitude; this is again similar to previous arguments. Toward consideration of the distribution, Kurt lists the distribution shapes, extrema, and outliers for the two groups. He then uses the distributions to explain the differences in center. He writes,

> The general length of sentences seems to be higher for group 2, but on the other hand group 1 has a lot of outliers. So according to the distribution group 2 has the higher length of sentences but in group 1 there are some individual cases with very high data points. This supports the claim of the researchers.

This is another combination of center and distribution, this time to point out that, despite having outliers on the high side, the center for group 1 is lower than for group 2.

There does not appear to be a trend in Kurt’s discussion of center, spread, and distribution over time. He correctly states and compares as appropriate the centers and spreads. He uses distribution primarily in conjunction with center, to choose measures of center to report and to explain graphically what the numerical representations show. This combination shows a deep understanding of each of these concepts.

5.3.2.2 HYPOTHESIS TESTING

In the third task, Kurt correctly states the hypotheses, using the symbol μ. His use of symbols rather than words is consistent with the example problems in class, which were written with the symbol μ, always accompanied by an oral description in the words of the problem. Kurt correctly states the p-value and the conclusion of the test as failing to reject the null hypothesis, but then asserts that the company does not profit from the warranties, a departure from the previous statement, which indicates an inconclusive result to the hypothesis test. Finally, Kurt states the effect size, making the error of not taking the absolute value. He interprets the effect size as small but does not directly link it to the result of the hypothesis test. One element Kurt does not include in his argument is the evaluation of assumptions to conduct the t-test. The sample size is large, which is sufficient to conduct the t-test. At the time of grading, the teacher-researcher interpreted the statement “Also the sample size is large enough to lead to reliable results” as an evaluation of assumptions, though upon further analysis, it instead appears to be related to previous statements that small sample sizes cause large variation and lack of reliability. Hypothesis testing in the third task is a weak area of Kurt’s arguments. He does not evaluate assumptions, makes errors in interpretation and effect size, and interprets the effect size but not as fully as possible.

In the fourth task, Kurt is much improved in carrying out the hypothesis test. He first evaluates assumptions, then explains the role of the t-test in the scenario: “The t-test helps us to investigate whether or not there is a difference in the judgment of different races because of potential biases.” He correctly states the hypotheses, p-value, conclusion, and effect size. His interpretation of the effect size, “It means that even though the hypothesis test came out to identify a difference between both groups, this difference can be considered to be small″ successfully synthesizes the effect size with the result of the hypothesis test, adding magnitude of the description of the difference between groups that has already been confirmed by the hypothesis test. Kurt shows clear growth in carrying out hypothesis tests between the third and fourth tasks. Unlike in the third task, in the fourth task he evaluates assumptions, calculates effect size correctly and interprets it fully, and gives the correct conclusion to the hypothesis test. He even adds a sentence explaining what the t-test accomplishes.

5.3.2.3 CONTEXT IN ARTICULATING RESULTS

In the second task, Kurt’s first attempt to incorporate context into articulation of results is in the statement, “The usage of the unit miles per hour makes perfect sense for the considered scenario.” This is the common mistake of interpreting the “uses units where appropriate” criteria on the required list of elements to include in a statistical argument as an instruction to evaluate whether the units are appropriate. Despite this misunderstanding, he correctly uses the units, miles per hour, in reporting means and outliers. He compares the means in context to one another, saying “neutral colored and black cars have slower average speeds than red colored cars.” He also points out that the mean speed for red cars is 16 mph over the speed limit and calls it “remarkable”, an indication he assesses the mean speed for red cars as high. He compares standard deviations in context, concluding there is “no remarkable variation in the collected average speed records of red and black cars” but that the standard deviation for neutral cars is twice as high. He further links to context by explaining the higher standard deviation in the neutral group as being because it includes different colors which “make the possibility for a spread higher.”

In the third task, when Kurt reports the median, he links to the context by observing, “The median looks promising for the company because the amount of claims is less than the revenue the company makes.” When he reports the standard deviation as being large, he concludes it is an indication that some customers have claims that are higher than the price they initially paid. For the hypothesis test, he interprets the conclusion in context as the company not making a profit. He does not, however, interpret the effect size in context.

In the fourth task, Kurt’s first mention of context in articulation of results is to evaluate the units as making sense for the data, showing he misinterprets the “uses units where appropriate” criteria for statistical arguments; this misconception is present in the second and fourth tasks, but not the third. When Kurt reports the medians in this task, he indicates the median for non-Caucasians is higher than the median for Caucasians, which supports the hypothesis of a difference between groups. Referring to the standard deviations of the two groups, he observes they are similar, and uses the context to assess them as high: “8 months of sentence are a long time, so in terms of the study it means the data varies.” For the hypothesis test and effect size, Kurt interprets the conclusion as there being a difference between groups, but that it can be considered to be small.

5.3.3 MAKING INFERENCES

In this section, Kurt's use of inference is discussed. As in the first two case studies, inference is divided into two parts: evaluation of sampling procedures and answers to the research question.

5.3.3.1 EVALUATION OF SAMPLING PROCEDURES

In the first task, Kurt finds two problems with the sampling procedures. The first problem he identifies is sampling bias. He judges the sample, which he identifies as a convenience sample, to be biased in terms of socioeconomic status, state of residence, and choice of recreational activity. He then makes a general recommendation that these issues should be remedied so the sample "can be representative of the population." From this statement, he is obviously considering the population; additionally, he understands and clearly explains how the sample is not representative of the population. The second problem Kurt identifies in this task is the sample size, which he says is "not enough to represent every teenager in America, especially if the sample is biased to a certain degree." In addition to Kurt's concern about the sample size and its interaction with sampling bias, this statement also reveals he considers the population to be all teens in the U.S. Taken together, these observations show Kurt understands the importance of having a sample that is representative of the population, and incorporates his understanding effectively into the first task.

In the second task, Kurt identifies two main problems with sampling procedures. First, he assesses the sample sizes for the three car color groups. He writes, "The researchers try to prove a general statement. In order to do so a very large sample size is required.” He appears to be using the term *general statement* to indicate making an inference to the population. Next, he identifies the sample as one of convenience, and says there is not enough variety in the sample to prove the “general statement,” again using his term. He recommends the sample be collected from a variety of cities and speed zones, saying it would be a more representative sample. These are similar to the two problems Kurt identified in the first task, and, though he does not specify what he considers the population to be, his use of the term *general statement* and discussion of sampling bias show he is thinking about inference.

In the third task, Kurt evaluates the sampling procedure favorably on the basis of its ability to show whether there is a profit. In this statement, he seems to be saying the data collected aligns with the research question, unlike in the first task, in which the data collected is about suffrage whereas the research question is about politics. He continues discussing the sampling procedure, claiming, “Another good feature of the data collecting procedure is that the sample they chose is a random sample. The first 250 customers could be from any race, age, gender and so on.” Kurt’s apparent conception of a random sample is that there is no obvious demographic bias in the sampling procedure. This is actually a convenience sample, but one in which the sources of bias are less obvious, such as systematic differences in consumer behavior or quality of the refrigerators between the first products sold and later ones. Still, the sources of bias in this task are less obvious, so their omission from the discussion of sampling bias does not indicate worsening performance.

In the fourth task, Kurt’s ability to identify sampling bias is revealed not to have declined. He evaluates the sample as being not representative of the population in two ways: data was collected from only one type of crime and from only one state. He subsequently recommends sampling from more than one felony, and from more than one state or country. This shows he considers the population to be all felonies in the U.S. or possibly the world.

There does not appear to be much change in Kurt's evaluation of sampling procedures over the course of the teaching experiment. However, there is little room for growth since Kurt's performance on the first task is so strong. In every task, he evaluates the sampling procedures to determine if they are likely to yield a representative sample. Where he finds bias, he clearly explains the source of the bias, at times recommending remedies.

5.3.3.2 ANSWERS TO THE RESEARCH QUESTIONS

In the first task, Kurt ends his argument by restating the problems he identified with the sampling and data collection procedures. Despite these issues, he does attempt to answer the research question. He writes,

> Finally, according to the study the research question can be answered as follows: The majority of teenagers do not pay attention to local and national politics, but more data is required from a survey with a more sufficient sampling and data collecting procedure and also a sample which is representing the population before it can be determined.

He makes a clear inference that the population of teenagers do not pay attention to local and national politics, but cautions that the inference can only be made based on data from a study that is conducted with more rigor. This is an ideal synthesis of the argument; the research question is answered based on the data given, with limitations based on the problems with sampling and data collection.

In the second task, Kurt's answer to the research question is almost as strong as his answer in the first task. He completes the second task by summarizing his conclusions. He concludes,

> Taking everything into consideration the research question can be answered as follows: according to the collected data it can be stated that drivers of red and black cars are driving faster than drivers of neutral colored cars in average. Furthermore especially drivers of red cars are driving way above average speed, but the data collecting procedure must be adjusted to achieve satisfying and reliable results.

He answers the research question in the affirmative; i.e., drivers of red and black cars drive faster on average than drivers of neutral colored cars, and he characterizes the magnitude of the red cars' speed as being large compared to the speed limit. Though he does provide an answer to the research question, he indicates the conclusion is made cautiously because of the problems he describes with sampling and data collection procedures earlier in the argument; this is another example of a . When making conclusions, Kurt does not explicitly state whether he is describing the sample or the population. However, his previous reference to the "general statement" indicates that he is likely inferring to the population. Other than this slight lack of clarity, his answer to the research question is an ideal inference.

In the third task, Kurt goes even further in his answer to the research question than in the first two tasks. He writes,

> Taking everything into consideration it can be initially said that the answer for the research question cannot be answered with yes. Most of the investigation towards the data analysis shows that the amount of claims tends to be higher than the revenue made by the company by selling the warranty in the first place. The only fact that speaks in favor of the company's pricing is the center of the data collection. The spread and also the distribution show the the data is strongly varying, especially towards a higher amount. Outliers are only on the higher side of the box plot too. Finally, the hypothesis testing also measures that the average

> amount of claims is not below 175 dollars. There is no recommendation in order to improve the data collection procedure. The research was done fairly well.

Here he not only states his answer to the research question; he also supports it by listing the evidence that supports it. More impressive still, he lists evidence that is counter to his conclusion and shows that it is outweighed by the evidence in favor of it. Here there is no need to limit the conclusion to account for biased sampling procedures, since he concludes they are not problematic.

In the fourth task, Kurt again lists evidence both supporting and countering his answer to the research question. He writes,

> Finally it can be concluded that there is a difference between the two groups according to the findings. The center, spread and distribution do not really support the fact that there is a reasonable difference among the judgment of both groups. Only the hypothesis test supports the initial thesis and gives us reason to further investigate the connection between different races and their penalties for crimes.

A weakness in this answer to the research question is the claim that the center, spread, and distribution do not indicate a difference in groups; this is counter to his earlier conclusion that the centers support a difference between groups, and he does not explain the discrepancy. Still, stating evidence countering the conclusion is more advanced than expected for the assignment.

There does not appear to be substantial changes in answering the research question over the course of the teaching experiment. Kurt’s answers to the research questions are clear, concise, and consistently strong in synthesizing conclusions based on sampling bias.

5.3.4 CONCLUSION

Kurt’s strong understanding of the course content shows in every aspect of his statistical argumentation. His statistical language is clear and concise, which is especially impressive considering he is not a native English speaker. He incorporates context throughout every task, using it to inform conclusions about the quality of sampling methods, the results of statistical analysis, and the answers to research questions. He is aware enough of context to recognize when an aspect of the intended context is different from the more familiar culture in which he has spent most of his life. He combines numerical and graphical representations of center, using the mean and median to explain aspects of a variable’s distribution, and vice-versa. He gives clear answers to the research questions that incorporate sampling bias as limitations on the conclusions drawn. There are few trends to report; Kurt’s written arguments began solidly and remained so throughout the teaching experiment.

CHAPTER 6: DISCUSSION

In this chapter, the results of the study will be examined as they relate to the three elements of statistical arguments: linking to context, articulating results, and making inference. The study's research questions of how these three elements change over the course of the teaching experiment will be addressed. In addition to the scaffolding for each element, characteristics of students' arguments will be discussed.

6.1 LINKING TO CONTEXT

In the first chapter, linking to context was described as consisting of any of the following: 1) a basic description of the phenomenon being studied, 2) a statement of the value or interestingness of the study, 3) a discussion of how the results make sense of the phenomenon being studied, and 4) a consideration of the credibility of the results. In this section, scaffolding of linking to context will be discussed. Characteristics of the ways in which students link to context will also be discussed.

6.1.1 SCAFFOLDING OF CONTEXT

The first research question in this dissertation is, "How do introductory statistics students change the way they incorporate the context of a problem into their statistical arguments over the course of a semester as they complete tasks designed to scaffold their argumentation skills?" The working hypothesis was that over time, students would increasingly ground their statistical arguments in the context of the data. Toward answering this research question, it is important to review the changes in each of the three case studies.

In the first case study, Amber shows little change in her use of context. In the first task, rather than focusing on the context as a whole, Amber recommends changing minute details; one example is she suggests specifying the ages at which people are considered to be teens. In the second and third tasks, she still focuses on details rather than the overall contexts; in both tasks, she recommends minor changes in wording the research questions that to her add clarification but do not appear to be improvements in a substantive way. Where she does improve is in her use of context in articulation of results. In the beginning of the teaching experiment, she does not relate the results to the given scenario and is confused about what units are used in the problem. By the end of the teaching experiment, she uses units correctly and interprets the results in context.

By contrast to Amber, who struggles with the scenarios presented, Leah appears to be very comfortable with each of the intended contexts. In each task, her contexts are shown to be in close alignment with the given contexts. Her interactions with context are deep and complex; she is similarly thorough in her use of context in articulation of results, though in the third and fourth tasks she misinterprets the “uses units where appropriate” criteria as a call to evaluate of the appropriateness of the units. Her use of context is in-depth from the beginning of the teaching experiment, and it does not appear to change over time.

Similarly, Kurt’s use of context does not appear to change over the course of the teaching experiment. In every task, his context is closely aligned with the intended context. He starts each argument with a summary of the scenario, which exceeds the expectations in the assignment and is even an improvement over the sample arguments provided. In every task, he uses context to evaluate sampling and data collection procedures, and to assess credibility. He also uses links to context as he articulates the results of each study, consistently interpreting the results of descriptive analyses, hypothesis tests, and effect sizes to the scenario presented. Similar to Leah, he also misinterprets the "uses units where appropriate" criteria; in this case, the misconception shows in the second and fourth tasks, but not the third task.

The hypothesis that students would increasingly ground their arguments in context does not appear to be supported by these three case studies. The only exception is in Amber's use of context in articulation of results, which does show improvement over the course of the teaching experiment.

6.1.2 CHARACTERISTICS OF STUDENTS' CONTEXTS

Analysis of the case studies presented in this dissertation yields four distinct ways in which students link to context in their statistical arguments. This section contains descriptions of each these ways of linking to context, with supporting examples.

The first way students link to context is by summarizing the information given in the task. Kurt begins each of his arguments with a short summary of the scenario. Beginning an argument this way serves to ground it in context. It has the additional benefit of allowing the argument to be read without the assumption that the reader has first studied the information provided to students. Amber and Leah also include some information given in the task in their arguments, but their summaries are limited to repeating the research questions.

A second way students link to context is by using it to identify sampling bias. Amber, Leah, and Kurt each identify sampling bias in the first task, and use their knowledge of context to recommend alternative ways likely to yield a more representative sample. Amber suggests sampling from different locations such as libraries and food courts, spread throughout various cities, while Leah proposes sampling public high schools throughout the country, and Kurt recommends sampling from different social classes within different states across the country. In some cases, they use context to provide additional insight into why sampling bias might occur. For example, in the second task Leah points out that cities may have different traffic patterns and differing tolerances of speeding, explaining how the sampling bias might affect the study.

A third way students link to context is by using it to identify problems with data collection procedures. In the first task, for example, students find the survey question about suffrage confusing and not in alignment with the research question. Also in the first task, Kurt identifies motivation to lie in the survey since an older adult, who is perhaps seen as an authority figure, is the person collecting data. In the second task, Leah uses her knowledge of context to point out that police officers may be biased against certain color cars, and thus they should not be responsible for recording the speeds.

The fourth way students linked to context in their arguments is by evaluating credibility of the results. Two examples come from Leah’s arguments. In the first task, she examines the results of the study and not only finds them to be inconsistent with her expectations of the research question, she goes further and determines them to be consistent with how she imagines respondents might have misunderstood the question. Another example comes from the third task, in which she uses context to explain the high number of outliers. She writes,

> Some models are low end, which would more than likely have more problems and there would be price differences in the parts used which would affect the claims paid. Higher end models would probably have fewer problems but parts might be more expensive. This situation might explain the number of outliers.

Whereas in the first task she evaluated the credibility of the aggregated results, in the third task she is evaluating individual observations; both of these are ways students use context to evaluate credibility of results.

6.1.3 FAMILIARITY OF CONTEXTS

During the course of this study, the effect of students' familiarity with contexts developed into a common theme. For some aspects of statistical argumentation as defined in this study—especially assessment of credibility of results and identification of sampling bias—students need to be quite familiar with the context. Identifying sampling bias requires students to understand the dynamics of the context enough to recognize when part of the population is overrepresented or underrepresented in the sample. Assessing credibility of results requires even greater familiarity with the context, since it requires students to compare the results of the study to their general knowledge of the context. Amber, Leah, and Kurt each had different experiences of the contexts, which impacted their arguments.

Kurt's experiences with the contexts are quite different from those of the other students, since he is less familiar with the culture in the United States. In the interview associated with the first task, Kurt stated that he does not know whether teens in the

United States are familiar with the word suffrage because he has spent most of his life in another country. In this case, Kurt’s history affects his assessment of credibility and his identification of problems with sampling and data collection. Not every example is so obvious, however, and every student brings his or her own experiences to the tasks. In the third task, after the interview Amber mentioned that she had never purchased a warranty and did not know how they worked. As a young adult, she does not have direct experience with a warranty, making it difficult for her to understand and discuss it. By contrast, Leah had two kinds of experience with this context: she had purchased a refrigerator with a warranty and she had seen a company conducting internal research. She reported that her experiences helped her use the results of data analysis to make sense of the contexts. As was seen particularly in the third task, however, previous experience with a context can become a distraction if the student puts too much emphasis on one particular instance in his or her personal history.

6.2 ARTICULATING RESULTS

In Chapter 1, articulating results was described as the discussion of the calculations, charts, and graphs generated from the data analysis. It includes center, spread, and distribution. It also includes of some inferential results such as confidence intervals and hypothesis tests. In this section, scaffolding of articulating results will be presented. Characteristics of articulating results will also be discussed.

6.2.1 SCAFFOLDING OF ARTICULATING RESULTS

The second research question in this dissertation is, “How do introductory statistics students change the way they articulate the results of data analysis into their statistical arguments over the course of a semester as they complete tasks designed to scaffold their argumentation skills?" The working hypothesis was that students would become more adept at combining the results of various charts, graphs, numerical summary measures, and hypothesis tests to articulate an overall description of the data. Examining the three case studies will address this research question.

In the first case study, Amber shows mixed results over time in her articulation of results. There does not appear to be any change in her discussion of center. Her incorporation of spread in her written tasks does not change substantially, but in the interviews she shows marked improvement in her fluency of discussing spread. For both center and spread, she primarily uses the numerical summary measures, reporting mainly the means and standard deviations without comparison or mention of magnitude. She becomes better able to identify the outliers shown on the boxplots, but her discussion of them lacks clarity. Her discussion of hypothesis tests and associated evaluation of assumptions and effect size are correct but unclear, and change little in the written tasks, though she loses some of the fluency in discussing them in the interviews. Overall, she struggles to articulate results throughout the teaching experiment, with improvements in some areas and minor setbacks in other areas.

Leah is much more consistent in her articulation of results, showing less of the fluctuation across tasks. Her discussion of center, spread, and distribution are strong in every task and show little change. She reports the numeric summary measures for center and spread, comparing across groups where appropriate. She shows improvement in the area of reporting hypothesis testing; in the fourth task she evaluates assumptions in a more logical place, corrects a misconception about the conclusion of the hypothesis test, and better integrates the effect size into the conclusion of the test. These improvements in the hypothesis test are substantial, but they are the only changes noted in Leah's articulation of results.

Similar to Leah, Kurt's discussion of center, spread, and distribution remains much the same over the course of the teaching experiment. In each task, he combines the numeric summary measures for center and spread with information about the distribution from the histograms and boxplots, which shows deep understanding of the relationships among concepts. Kurt shows another similarity to Leah in that he shows marked improvement between the third and fourth tasks in discussing the hypothesis test; he adds the previously-omitted element of evaluating assumptions, corrects two errors, and better integrates the effect size with the result of the hypothesis test.

Taking the three case studies together, it is reasonable to conclude that there are improvements in some of the individual elements of articulation of results. For each of the three students, areas that are initially weak do seem to get stronger, and there are very few instances of worsening performance on tasks. This is particularly true of the hypothesis test, which for Leah and Kurt improved greatly. Center, spread, and distribution change little, though Amber shows markedly better fluency in discussing spread in the interviews. In general, their discussion of each individual element tends to improve over the course of the teaching experiment.

In addition to the progress of individual elements of articulation of results, a second type of scaffolding is observed. The design of the teaching experiment is that the statistical methods included in the tasks accumulate over time. In the first teaching episode, the results consist only of relative frequencies; center, spread, and distribution are added in the second teaching episode; a one-sample t-test is included in the third teaching episode; and an independent samples t-test is included in the fourth teaching episode. Scaffolding is demonstrated if students maintain the level of articulation of results from the previous tasks, while successfully including the new concept(s) of the current task. In this way, scaffolding is clearly demonstrated in this teaching experiment. In each task, each of the three students remained at the same, or better, quality of articulating results as in the previous tasks. Additionally, these students were successful at incorporating the new statistical concepts in each task. As was hoped at the outset of the teaching experiment, students were able to accommodate increasing statistical content into their statistical arguments.

6.2.2 CHARACTERISTICS OF LINKING CENTER AND SPREAD TO CONTEXT

The statements students make in discussion of center and spread can be categorized into three levels, which show increasing depth in interpreting the results in context. The first level is listing descriptive statistics, which consists of simply reporting the mean, median, standard deviation, and/or range. The second level is comparing, which may be to a constant value or to another group. The third level is assessing magnitude, subjectively determining the size of a measure of center or spread, or the size of a difference across groups.

At the most basic level, students listed the descriptive statistics from the SPSS output. They use units do not do any comparison or interpretations beyond stating the results. An example of this is from Amber’s second task. She writes, “The speed of neutral-colored vehicles has a mean of 65.9667 units. The speed of black-colored vehicles has a mean of 69.4348 units. The speed of red-colored vehicles has a mean of

76.4091 units." Here Amber states the values, attempting to use the units from the problem, and she describes the group in words, but she does not follow the statement with an interpretation of the results in context.

At the intermediate level, students accompany the descriptive statistics with a comparison. Two examples come from Leah's work. In the second task, she first states the values of all three medians, then follows with the statement, "Neutral colored cars have the lowest median mph with 67.00 and red colored cars have the highest median mph with 76.00." In this case, Leah points out the values of the groups with the highest and lowest medians, a comparison. A second example is from her fourth task: "The standard deviation for group 1 is 8.22424 months and for group 2, 7.10535 months. This data shows group 1 having a greater distance from the mean than group 2." In this passage she compares the two groups, but she also interprets the values as distance from the mean.

At the third level, the most advanced, students discuss center and spread in assessing magnitude. This includes comparison, but also includes a subjective assessment of the size of a difference. Two examples are seen in Kurt's work. In the second task, after stating and comparing the mean speeds of the three groups, he considers the largest mean in relation to the speed limit, an important value in this context. He writes, "Drivers of red colored cars drive around 16 mph over the speed limit which is remarkable." He goes beyond comparing the mean to the speed limit; he calculates the difference and finds it to be "remarkable" in size. This assessment of the difference as large is a deeper level of analysis than comparing the means. A second example comes from his fourth task, where he does a similar assessment of magnitude for standard deviation:

> The standard deviation of group 1 is 8.2242 months of sentence received and the standard deviation for group 2 is 7.10535. The spread of the data is similar among both groups, but it can also considered to be high. 8 months of sentence are a long time, so in terms of the study it means that the data varies.

In this context he does not have a constant value for comparison, but he considers the size of the standard deviation in context as a prison sentence. When considered in that context, he determines the standard deviation is quite large.

6.2.3 KURT'S USE OF VISUAL REPRESENTATIONS OF CENTER AND SPREAD

In his tasks, Kurt demonstrates an unusually advanced understanding of how distribution relates to center and spread. The second, third, and fourth tasks in this teaching experiment include boxplots and histograms of the data as part of the output. Boxplots and histograms are graphical (or visual) representations of the data, and they are the primary tools for characterizing distribution in the introductory statistics curriculum. When asked to discuss distribution, students refer to the boxplots and histograms, and when asked to discuss center and spread, most students use the numeric summary measures. Kurt, however, combines his discussion of distribution with center and spread. He does this in two distinct ways. First, he uses the boxplots and histograms to characterize center and spread visually. They serve to reinforce the numeric measures of center and spread he reports elsewhere in his arguments. Second, he uses the boxplots and histograms to identify characteristics of the distribution that explain the numeric summary measures.

There are two instances in which Kurt characterizes center and spread graphically. In the first instance, he uses boxplots to compare center and spread simultaneously. He writes in the second task,

> The box plot of neutral colored cars has the biggest range and the lowest median. The box plots of the red and black colored cars are very similar in their range, just the median of the red colored cars is higher than the median of black colored cars.

Here he is referring to the numeric summary measures, but he is considering them visually based on the boxplots. In the second instance, also from the second task, he relates the boxplot to the center: “The fact that there is an outlier at the data of black colored cars which is below the normal range underlines the fact that black colored cars are driven slower than red colored cars in average.” In this case he uses the low outlier in the black cars to “underline,” or reiterate, the lower mean. Each of these instances show Kurt using graphical methods of looking at center and spread to supplement the numeric methods.

There are also three instances in which Kurt uses features of the distribution to explain the dynamics that caused the measures of center and spread to occur the way they did. The first instance comes from the second task. He writes, “The fact that the graphs of the data of red and black colored cars is left-skewed means that there are more higher data points. So red and black cars are driven faster in average.” Here he uses the shape of the distributions of red and black cars to explain why these two groups had higher means than the neutral cars. Similarly, in the third task, Kurt uses the distribution to explain both the center and spread. He writes, “The histogram is strongly right skewed which indicates that low data points are most common. This explains why the median is fairly low even though there are a lot of higher outliers and a fairly high standard deviation.” In this case, the skewness explains the low median and large standard deviation. The third instance also comes from the third task. He describes the standard deviation as large, then explains, “It basically means that there are probably a lot of outliers and that there are a lot of customers who have claims which are higher than the price they initially paid.” Though he phrases it as conjecture, in this statement he uses outliers and other high values in the data to explain a large standard deviation.

Kurt’s use of boxplots and histograms to characterize center and spread visually and to explain the magnitude of the values of the numeric summary measures show a deep understanding of these concepts. The way he combines these visual and numeric representations can become an ideal for students in introductory statistics.

6.2.4 STUDENTS’ DISCUSSION OF HYPOTHESIS TESTS

The next feature of students’ articulation of results is the way they discuss the hypothesis test results and effect sizes. In this section, two aspects are explored: students’ reporting of the inconclusive result in the third task, and their incorporation of effect size into the conclusion of the results in the fourth task.

The third task requires students to test whether the mean cost to replace or repair refrigerators exceeds $175; the p-value dictates an inconclusive result, which each student has a slightly different way of interpreting. Students’ difficulty understanding inconclusive results is well-documented (Batanero, 1994). In particular, it is challenging for them to distinguish between an inconclusive result (failing to reject the null hypothesis) and the stronger, incorrect conclusion of confirming the null hypothesis. It is worthwhile to examine the conceptions of the students in this study.

In Amber’s statement of the conclusion of the hypothesis test, it is not clear if she understands the conclusion. She writes, “There is not sufficient evidence to conclude that the amount of claims the company must pay to repair or replace the refrigerators is less than $175 on average, and they will make a profit.” The way the sentence is written, it is impossible to determine if she means to assert that the company will make a profit, or if she intends “they will make a profit” to be part of what there is not sufficient evidence to demonstrate. In the interview, she interpreted the result as the latter. This shows she did not, at least in the interview, make the error of conflating an inconclusive result with confirmation of the null hypothesis.

In Leah’s argument, she states her conclusion at two different places, and the two conclusions are not the same. She first concludes there is “insufficient evidence that $\mu <$ 175 dollars claimed,” showing she interprets the result as inconclusive. However, later in the argument when she answers the research question, she determines “the data proved that the company did not spend on average less than 175.00 dollars on a claim.” This is a stronger conclusion than she previously reported; in this one, she is asserting that the test confirms the null hypothesis. When asked about the conclusion in the interview, she reported it the way she first does in her argument, that it cannot be determined the company made a profit. Though in the second mention of the conclusion in her argument she overstates the conclusion, she does not overstate it in the interview. It appears she experiences some confusion during the writing of the task that has been resolved by the time of the interview.

Kurt’s phrasing is similar to Leah’s. His conclusion consists of two conflicting sentences, “There is no sufficient evidence that the average amount of claims that has to be covered by the company is less than 175 dollars. So there would not be a profit for the company." In the first sentence, he states the result as inconclusive, but then in his interpretation he confirms the null hypothesis. Like Leah and Amber, in the interview he has no trouble stating the result of the hypothesis test as failing to reject the null hypothesis. When asked during the interview if the conclusion shows confirmation that the company does not make a profit, Kurt confidently answered that it does not. It seems Kurt also experiences some confusion during the writing which is no longer present by the time of the interview.

Each of the three students experienced some difficulty with the conclusion of the hypothesis test in the third task. For Amber, her conclusion lacks clarity, while Leah and Kurt overstate the conclusion. In the interview associated with the third task, all three students correctly stated the conclusion and did not appear to struggle. When students completed the written task, they were still fairly new at conducting hypothesis tests. By the time of the interview, they had more practice with hypothesis testing and had refined their understanding.

The other characteristic of interest in students' discussion of hypothesis tests is the way they report effect size. Two aspects of their reporting of effect size are examined. The first is whether students synthesize effect size with the conclusion of a hypothesis test. For statistically significant tests, effect size provides additional evidence that either strengthens or limits the conclusions, what Abelson called *tics* and *buts*; it was hoped that students would discuss effect size in those terms. The second aspect of their reporting of effect size is whether they, as Cohen (1988) recommends, treat the guidelines as rough suggestions that are dependent on context and the subjective judgment of the researcher rather than rigid rules of interpretation. Since the hypothesis test in the fourth task yields a statistically significant result, students' discussion of the accompanying effect size, calculated as Cohen's *d*, reveals their understanding of these two aspects of effect size.

In her fourth task, Amber's statements reporting the hypothesis test and effect size are correct. However, she does not link them. She writes,

> There is statistically significant enough evidence to conclude that Caucasian offenders will get more lenient sentences than any other group. The data is a small effect of .2698; the guidelines are .2, .5, and .8. A small effect size means that there is hardly any difference between the data we have been given for this research.

In this passage, effect size is stated separately rather than synthesized with the conclusion to the hypothesis test. Amber's interpretation reveals she considers the guidelines to be fixed and rigid. Her understanding appears to be at the emergent level, where she correctly calculates and states the effect size, but does not demonstrate the more subtle aspects of subjectivity and synthesis.

Leah's interpretation is at a slightly deeper level. She writes, "Our effect size is .2699, which is small and is not practically significant. … Even though Non-Caucasians have a higher average mean for sentences received, those sentences were not high enough over Caucasians mean to practically support the study." Leah synthesizes the effect size into the conclusion of the hypothesis test, stating it as a limitation on the statistically significant result. She does not specify how she is applying the guidelines.

Kurt's interpretation shows both synthesis and an understanding of the roughness in the guidelines for effect size. He writes,

> There is sufficient evidence to conclude that Non-Caucasian people do get longer sentences for their crimes of burglary. The effect size, calculated through

> Cohen’s d, came out to be 0.2698, which is on the smaller end. It means that even though the hypothesis test came out to identify a difference between both groups, this difference can be considered to be small.

Like Leah, he uses effect size to limit the result of the hypothesis test. Additionally, his use of the phrase “the smaller end” shows he understands the approximate nature of the guidelines. This is an advanced level of interpretation of effect size.

6.3 MAKING INFERENCES

In Chapter 1, making inferences was described as including all aspects of the statistical argument related to generalizing from the sample to the population. It includes an assessment of whether the data collection procedures are likely to yield a sample that is representative of the population. It also includes the impact of sampling bias on generalizing to the population. In this section, student work in making inferences is discussed. Scaffolding and characteristics of making inferences are also presented.

6.3.1 SCAFFOLDING OF INFERENCE

The third research question in this dissertation is, “How do introductory statistics students change the way they make inferences in their statistical arguments over the course of a semester as they complete tasks designed to scaffold their argumentation skills?” The working hypothesis was that students would better identify bias in sampling procedures and adjust the claims and conclusions made in their statistical arguments accordingly.

In the case of Amber, there is no notable trend in her ability to identify sampling bias; she consistently notes sampling bias and finds the sample sizes to be too small. However, she shows growth in answering the research questions given in the tasks. In the first two tasks, she does not answer the research questions in writing, though when prompted she has no trouble answering them. In the third task, she answers the research question, but does not use the sampling bias to limit the conclusion. In the fourth task, she shows improvement by synthesizing, though vaguely, the sampling bias into the answer to the research question. This is remarkable progress.

Leah's progress over time closely mirrors Amber's. Leah is very thorough in identifying sampling bias in each task, and there does not appear to be change across time. She does, however, improve in answering the research questions. Like Amber, in the first two tasks, she does not attempt to answer the research questions. She focuses instead in these two tasks on evaluating the quality of the information provided. In the third task, she answers the research question by reiterating the result of the hypothesis test and effect size, without synthesizing the sampling bias into the answer. In the fourth task, she improves by connecting the sampling bias to the conclusion. It is not fully synthesized, but it is connected nonetheless. Over the course of the teaching experiment, both Amber and Leah improve from not attempting to answer the research questions in the first two tasks to answering the research questions including basic syntheses of sampling bias. For both Amber and Leah, the hypothesis tests in the third and fourth tasks appear to facilitate their answers.

Like Leah and Amber, Kurt is consistently strong in identifying sampling bias; in every task, he evaluates the sampling and data collection procedures, clearly explaining any sampling bias he finds. In the first task, he answers the research question clearly, and the cautionary note he includes with the answer shows synthesis. With the first task he achieves the desired outcome. Unlike them, his inferences in each of the four tasks are direct and include synthesis of sampling bias.

Considering these results together, it does appear that inference is being scaffolded. Each of the students were strong in identifying sampling bias throughout the teaching experiment and did change; however, the answers to the research question improve over time. In particular, the addition of a hypothesis test beginning in the third task is a turning point, prompting Amber and Leah to include an inference, in writing, for the first time. Since hypothesis testing is inherently inferential, it is not surprising that it helps bridge from articulation of results to inference. For Amber and Leah, the conclusion of the hypothesis test served as a reminder that the statistical argument should in the end answer the research question.

6.3.2 CHARACTERISTICS OF STUDENTS' INFERENCES

One of the more advanced aspects of statistical argumentation is synthesizing sampling bias into the answer to the research question in a given task, using it to limit the conclusion. In this section, the different ways Amber, Leah, and Kurt address the research questions are discussed, with particular emphasis on the depth with which they consider the sampling bias in their conclusion.

At the most emergent level of inference is when a student does not make an inference. In their first two tasks, both Amber and Leah conclude their arguments with summaries of the flaws they found with sampling and data collection. An example is from Leah's first task, "As we can see this study has numerous flaws, but with proper sampling and data collection procedures, the study could be sound and useful to the public." She does conclude the argument, but her conclusion serves to evaluate the information provided rather than address the research question.

A second level of inference is when a student answers the research question but does not synthesize flaws of sampling and data collection into the answer. Amber does this in her third task. Her conclusion is,

> As I stated before, if the study rephrases the research question to be able to include only the specific brand of refrigerators, the study would be more accurate and provide better data for the research. The assumption to run the t-test has been met, however the study did not provide sufficient amount of evidence to conclude that the company will make a profit if the amount of claims the company must pay to repair or replace the refrigerators is less than $175.

In her conclusion, Amber reiterates the flaws she identified in the study. She also states her answer to the research question, but she does not use the flaws to limit the inference.

A third level of inference occurs when a student attempts to synthesizes sampling bias into the answer to the research question, but the synthesis is vague or unclear. Amber's conclusion in her fourth task does this. "Even though I feel like the research sample and data collection procedure could change. The research still included enough evidence to conclude that Caucasian offenders will get more lenient sentencing than any other group." Her use of the words "even though" shows she is connecting the flaws to the conclusion, using them to detract from her conclusion, but she does not fully explain her reasoning.

A fourth level of inference consists of fully synthesizing the sampling bias into the answer to the research question. Among the three students in this study, only Kurt achieves this level of inference. An example is shown in his second task, where he concludes,

> According to the collected data it can be stated that drivers of red and black cars are driving faster than drivers of neutral colored cars in average. Furthermore especially drivers of red cars are driving way about average speed, but the data collecting procedure must be adjusted to achieve satisfying and reliable results.

Here answers the research question, specifying that he is speaking for the data that was actually collected. He then uses the flaws in data collection to mitigate his answer. His conclusion may have been improved if he had specified that the conclusions could be inferred only to the one city and speed limit which all the data was collected. Nonetheless, it represents a near ideal level of inference in these tasks.

CHAPTER 7: REFLECTIONS AND RECOMMENDATIONS

In this chapter, significance of the study will be discussed. Next, feedback from the students and reflections from the teacher-researcher will be presented. Third, recommendations for future instruction of statistical argumentation and for future research will be discussed. Lastly, final thoughts consisting of factors to consider for other instructors who are planning to engage in statistical argumentation will be presented.

7.1 SIGNIFICANCE OF THE STUDY

This study contributes to the literature in statistics education in several ways. First, the development of the definition and heuristics of statistical argumentation offers specificity to a type of discourse common in academic and workplace settings. Articulating the heuristics lays the groundwork for further analysis of statistical argumentation in these everyday contexts.

Second, this study shows that with scaffolding, postsecondary students at the introductory level can engage in statistical argumentation. This is in contrast to Abelson's (1995) book on statistical argument, which Abelson said was intended for more advanced students, those on at least the third pass-through of statistical concepts. Students' success engaging in statistical argumentation is possibly the most important result of the study.

Third, this study points to some benefits of statistical argumentation. For Leah and Amber, who struggled with calculations, statistical argumentation provided an alternate way of assessing their understanding of course concepts. It was effective as a means of assessment, because in all three case studies, students' arguments provided information about misconceptions that would not have been revealed through traditional calculations. For Leah, it allowed her to leverage her life experience in the statistics classroom, which indicates the particular value of statistical argumentation for adult learners.

Fourth, this teaching experiment shows how statistical argumentation can be incorporated into the introductory statistics curriculum while still meeting existing course objectives. Prior to this dissertation, only three studies were located in which efforts were made to facilitate statistical argumentation. Of those, only one (Derry, Levin, Osana, Jones, & Peterson (2000) was at the postsecondary level, and it took place in an enrichment course rather than in an introductory statistics course.

Finally, this study points to a new kind of scaffolding of statistical argumentation. Hudson (2010) identified two kinds of scaffolding: contextual scaffolding, in which the teacher helps students understand the problem and how the ideas in one context can be applied to a problem in another context, and social scaffolding, in which the teacher helps students learn to engage in statistical argumentation through one-on-one interaction. In this study, both contextual and social scaffolding were employed. However, a third kind was essential: structural scaffolding, in which tasks are given to students in an order that builds their statistical argumentation. In this study, structural scaffolding occurred in the way arguments were based on increasingly advanced statistical content and students created arguments in small groups before working on their own.

7.2 REFLECTIONS

At the end of the interview associated with the fourth task, students were asked to provide feedback on the assignments. They were encouraged to be honest and that negative feedback would not impact their grades or hurt the teacher-researcher’s feelings. Despite these assurances, there is no way to determine whether their responses were affected by the lack of anonymity. This feedback is reported here as student reflections on the teaching experiment. This section also includes teacher-researcher reflections.

7.2.1 STUDENT FEEDBACK

Students were first asked a general question, “Do you think these assignments have been beneficial?” All three students responded in the affirmative. Amber said they helped her study; she indicated they reinforced the course concepts being studied. Kurt explained that he liked seeing how the material being studied in class would be applied in the real world. Leah said she found them thought-provoking and she loved doing the assignments. She said she’s an analytical thinker but is intimidated by numbers. In general, she feels she is “not good at math” and she was panicked by the time constraints on in-class tests and quizzes. She likes that these assignments allowed her to show her understanding in a less computational intensive way and that she could complete them in a more comfortable environment. When asked whether she thought the experiences she brings to these tasks as an adult student helped, she said she thought they did help somewhat. She added that the particular experience of being a paralegal made a big difference, since part of her job as a paralegal was to help in analyzing cases and constructing arguments.

Students were next asked what they thought were some negatives about the assignments. Kurt said he could not think of any. Amber also said she could not think of any, except that the second task contained several problems with sampling and data collection. Leah expressed frustration with the group tasks, citing poor communication among group members.

In the final part of the feedback, students were told, “There were several ways in these assignments I tried to build your statistical argumentation over the course of the semester. I’d like to go over a list of those and have you tell me whether you think it was helpful or not.” The strategies included the example arguments, list of required elements, group assignments, feedback on group assignments, face-to-face interviews associated with the tasks, and structure of the assignments as increasing statistical content throughout the semester. Though not interviewed together, the three students were remarkably similar in their responses.

The students all agreed that the only strategy they did not enjoy was working in groups. Kurt found it difficult for group members to agree on phrasing in the arguments, while Leah and Amber both felt some of the group members did not contribute enough or communicate in a timely manner. Each student was asked a follow-up question, whether they thought the group members helped each other so that in the next assignment they would write better individual arguments. All agreed that yes, despite their frustrations with group work, they did feel their individual arguments benefitted from having a group argument prior.

For the other strategies to scaffold statistical argumentation, all three of the students agreed that the feedback on group assignments was helpful; they reported that they made sure any suggestions on the group assignments were incorporated into the individual assignments. They liked having the list of required elements, and they all used it as a checklist to make sure they had not omitted any elements. Similarly, they all agreed that the structure of the assignments, with the increasing amounts of statistical content, was helpful. Kurt and Leah each reported they felt more comfortable with statistical argumentation as they got more experience with throughout the semester. For the example arguments, Leah and Amber found them to be beneficial. Leah felt they helped make the statistical language not seem foreign, and Amber said without them she would have been confused and unsure of the expectations. Kurt, on the other hand, said that though the example argument in the first teaching episode was very important, he did not need them in the later tasks.

Of all the strategies in this study designed to scaffold statistical argumentation, Amber, Leah, and Kurt were most enthusiastic about the interviews. Kurt felt they were more beneficial than any other aspect of the assignments; he said it was a way to see if he really knew the material because making the arguments orally was more challenging than making them in writing. Leah reported that she enjoyed them because she prefers face-to-face communication, though she found answering questions a little stressful. Amber said she liked them because she could ask questions as they came up during the argument, giving her a chance to clarify concepts she did not understand.

7.2.2 TEACHER-RESEARCHER REFLECTIONS

An unexpected result during data collection is that more students chose the reflections by interview than was anticipated. The total number of interviews for the four tasks was 27, 21, 22, and 19 students, respectively. This was a challenge because on many occasions, there were several student interviews in a row, leaving the teacher-researcher without time to review students' prior work immediately prior to the meeting or record detailed notes after the meeting. As a result, some of the work modeling student understanding was less iterative between meetings than would have been in an ideal teaching experiment. However, in this study the interviews were secondary data to the written tasks, used primarily to clarify students' written work, so the impact on this study is minimal. A positive aspect is that the students seemed to enjoy and benefit from it, as shown by their repeated choice of interview as the method of reflection, and their feedback as reported in the previous section. The interviews were also personally rewarding. One example is that Amber struggled with spread and was memorably excited when she knew how to discuss it in the last interview. Though very time-consuming, these kinds of interactions made the interviews enjoyable.

In retrospect, one likely reason students do not appear to show improvement in the context criteria is found in the structure of the tasks. The first teaching episode is designed to introduce students to the norms of statistical argumentation while keeping the statistical content minimal, meaning it focuses primarily on context and inference. Additionally, the context element does not increase in difficulty in the remaining three teaching episodes. Thus, the criteria of context is primarily scaffolded in the first teaching episode. Thereafter, growth may be expected in response to feedback on the tasks and in the interviews, or by learning from classmates' use of context on the group tasks.

A small but significant observation from Leah and Kurt points to another possible reason: the length of time between tasks. Kurt's misconception about the meaning of the "states units where appropriate" criteria on the list of elements to include is shown in the second and fourth tasks but not the third. Leah has the same misconception in the third and fourth tasks. This criteria was clarified in class (both Kurt and Leah were present) and in the interviews with both students. This is a fairly straightforward clarification, requiring students to understand only that they need to use the units when stating descriptive statistics rather than evaluate whether the units are appropriate for the scenario; essentially, it is a matter of recall. Leah did not retain the information between the third and fourth tasks. Kurt retained it between the second and third tasks, but did not retain it for the fourth task. There were two weeks between the second and third tasks, and another two weeks between the third and fourth tasks. From this data, it appears that over time, this misconception reemerged.

7.3 RECOMMENDATIONS

7.3.1 RECOMMENDATIONS FOR INSTRUCTION

The results of this study point to the importance of two characteristics of contexts in statistical argumentation tasks. The first is that the context should be familiar. For students to be able to identify sampling bias and assess the credibility of results, they must have a general knowledge of the context. Much of the emphasis in the statistics education literature—Libman (2010) and Dierdorp, Bakker, Eijkelhof, & van Maanen

(2011) are two examples—is on authentic contexts, the use of real-world data for statistical analysis. Many of these authentic contexts may be from areas such as science or industry, which may be understandable but not familiar to students. Contexts that are authentic are still ideal, but for purposes of statistical argumentation they must also be familiar.

Another characteristic of the contexts is that they should be interesting, but minimally controversial. In this study, the fourth task was about racial bias in sentencing for felonies. While this is an important question that can be addressed statistically, for purposes of these tasks it is too sensitive, and because of current events at the time this task was completed it was even more so. The sensitivity of the fourth context meant that pressing students to reflect on their experience with and knowledge of it ran the risk of causing emotional distress. In the interest of ethics as well as keeping the interviews on track, the discussion was limited to asking each student what he or she knows about the context, whether he or she thinks there is racial bias in sentencing; there was no follow-up. This limited both the ability to facilitate interaction with the context and the ability to find out about each student's context.

At the outset of this study, a distinction was made between learning to argue and arguing to learn. It was said that the primary goal of introducing statistical argumentation was to help students learn to argue, i.e. for them to learn how to communicate the results of statistical analysis. By contrast, arguing to learn, the process of employing argumentation as a way to reinforce course concepts, was considered a secondary goal. Students were not asked about this in the feedback portion of the last interview, but both Kurt and Amber volunteered that they felt doing these tasks help them understand and learn to apply the statistical methods learned in the course. The results of this study show success in learning to argue, and the feedback from students shows they also believe they are arguing to learn. This goal should also be considered in future implementations of statistical argumentation. For example, if time between tasks hinders scaffolding, it may be possible to rearrange the course so that all the statistical argumentation tasks are completed in a separate unit at the end. While the reduced amount of time between tasks might better serve the goal of learning to argue, it would separate the statistical arguments from the learning of the concepts; thus it would hinder the goal of arguing to learn. The schedule in the current classroom teaching experiment attempted to balance these two goals.

A final recommendation for teaching statistical argumentation is that when outliers are present in data analysis, it is valuable to ask them to evaluate whether the outliers are more likely to be errors or legitimate observations. In this teaching experiment, it proved to be a valuable exercise regarding context. It is an opportunity for students to interact with individual observations, the only such opportunity in these tasks. It also allows the teacher-researcher to determine whether students know what each observation represents (the speed a particular car is traveling, the cost of repairs for an individual refrigerator, or the number of months in an offender’s prison sentence). For example, in Amber’s third task, asking her about the outliers revealed that she did not understand what the values represent, which provided an opportunity for clarification. Another benefit is that it encourages a student to reflect on his or her context—what is typical or believable in the student’s experience. At the same time, it provides insight into the student’s context. In two ways, then, asking students to evaluate whether outliers are errors or legitimate observations serves the dual purpose of facilitating deeper interaction with the context while simultaneously allowing the teacher-researcher to assess students' knowledge.

7.3.2 RECOMMENDATIONS FOR FUTURE RESEARCH

As statistical argumentation is a new concept in statistics education, there is much to be learned about it. This study was conducted in sections of university-level introductory statistics containing at most 25 students each section. Larger classes would require adjustments since instructors cannot spend as much time with each student as was required for this study. Online sections would also require adjustments. Students at different levels—middle and secondary grades, introductory postsecondary, and advanced or graduate-level postsecondary—will have differing needs. Further study might focus on variations of these environmental factors.

Effective instructional techniques for fostering statistical argumentation would also be a rich area of research. These might include varieties of the types of tasks assigned, ways of introducing tasks, and conditions of feedback. This study yielded insights into levels of depth of analysis in some areas. It would be useful to build on those to develop and test a rubric for assessing statistical arguments.

Two aspects of articulating results arose that require further research. The first is students' understanding of the relationship between visual and numeric characterizations of center and spread. The second is students' understanding of effect size, and the language used to communicate results of effect size in conjunction with hypothesis test results. Since it is still not common for effect size to be included in introductory statistics courses, it would be valuable to determine if discussion of effect size gives students a more nuanced understanding of hypothesis testing.

Finally, there is a dearth of research on adult learners of mathematics and statistics. As more and more postsecondary students are adult learners, serving this population first requires understanding their particular needs and strengths. Research should also focus on identifying pedagogy for adult learners (what Knowles (1973) called andragogy) in mathematics and statistics that is effective for them.

7.4 FINAL THOUGHTS

The final thoughts in this dissertation consist of factors for instructors to consider in embarking on a plan to incorporate statistical argumentation in their own classes. There are a few things to consider. First, it is very time-consuming. It takes a substantial amount of class time to present the norms of statistical argumentation and allow students to work in groups. In this study, to free up class time the number of tests were decreased from four to two. Tests then consisted only of problems; understanding of larger concepts did not need to be on the tests since they were being assessed in the argumentation activities. It also takes a substantial amount of time outside of class. There were eight tasks to grade, and students need detailed feedback in order to improve. Additionally, students need the feedback on each task (both the group and individual tasks) prior to starting work on the next task, so grading needs to be timely. On top of the time for grading, meeting with students for interviews adds to the time commitment. Institutional and departmental characteristics such as typical class size, schedule of class meetings, use of department-wide exams, and availability of assistance with grading may necessitate some variations to the program of statistical argumentation presented in this dissertation.

A second thing to consider is the importance of creating a class culture that supports argumentation in general and statistical argumentation in particular. This became clear with successive iterations of the program of statistical argumentation. It is not enough to engage in statistical argumentation only during the discussion of tasks. Beginning on the first day of class, students can be introduced to the ideas of statistical inference, center, spread, and distribution, as well as the importance of communicating the results of statistics. Nearly every day of the course presents opportunities to incorporate elements of statistical argumentation. When measures of center are introduced, for example, students can be given a scenario with medians of several groups, and they can be guided to discuss the results by comparing the medians, evaluating the magnitude of the differences in the medians, and determining what they reveal about the context. Developing a culture that supports statistical argumentation requires commitment, but it also helps focus the course on statistical literacy and reasoning, which is in keeping with the recommendations discussed in Chapter 1.

A final consideration is that there are rewards of a program such as this both to students and instructors. It is a way of assessing students' understanding, one that allows students whose strengths are not often leveraged to excel in introductory statistics. It gives adult students a way to bring their life experiences to their statistics course. Students tend to see the value of this kind of work, and thus are engaged and motivated. The individual reflections bring students for one-on-one conversation who might not do so otherwise, and talking through the statistical arguments is another way to reinforce the course concepts. Seeing and celebrating students’ successes is also rewarding. These benefits make the commitment to statistical argumentation in introductory statistics classes worthwhile.

REFERENCES

Abelson, R.P. (1995). *Statistics as principled argument.* Lawrence Erlbaum Associates Publishers: Hillsdale, NJ.

Batanero, C., Godino, J., Vallecillos, A., Green, D., & Holmes, P. (1994). Errors and difficulties in understanding elementary statistical concepts. *International Journal of Mathematical Education in Science and Technology,25*(4), 527-547.

Begg, A. (1997). Some emerging influences underpinning assessment in statistics. In I. Gal, & J. Garfield (Eds.), *The assessment challenge in statistics education*. Amsterdam: IOS Press.

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3-16). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Cho, K. L., & Jonassen, D. H. (2002). The effects of argumentation scaffolds on argumentation and problem solving. *Educational Technology Research and Development*, *50*5-522.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *American Mathematical Monthly*, 104(9), 801-823.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates Publishers: Hillsdale, NJ.

Derry, S. J., Levin, J. R., Osana, H. P., Jones, M. S., & Peterson, M. (2000). Fostering students' statistical and scientific thinking: Lessons learned from an innovative college course. *American Educational Research Journal,37*(3), 747.

Diaz, R. M., Neal, C. J., & Amaya-Williams, M. (1990). The social origins of self-regulation. In L. C. Moll (Ed.), *Vygotsky and education: Instructional implications and applications of sociohistorical psychology* (pp. 127-155).

Dierdorp, A., Bakker, A., van Maanen, J., & Eijkelhof, H. (2012). Supporting students to develop concepts underlying sampling and to shuttle between contextual and statistical spheres. Paper presented at the 12th International Congress on Mathematics Education, July, 2012, Seoul.

Forman, E.A., Larreamendy-Joerns, J., Stein, M.K., & Brown, C.A. (1998). “You’re going to want to find out which and prove it”: Collective argumentation in a mathematics classroom. *Learning and instruction, 8*(6), 527-548.

Hitchcock, D. (2002). The practice of argumentative discussion. *Argumentation, 16*(3), 287.

Hite, S. (1987). *Women and love: A cultural revolution in progress*. Knopf: New York.

Hogan, K., & Pressley, M. (1997). Scaffolding scientific competencies within classroom communities of inquiry. In K. Hogan & M. Pressley (Eds.), *Scaffolding student learning: Instructional approaches and issues* (pp. 74-107). Brookline Books: Cambridge, MA.

Hudson, R.A. (2010). *Statistical argumentation in project-based learning environments* (Doctoral dissertation). Retrieved from http://proquest.umi.com.librarylink.uncc.edu/.

Huff, D. & Geis, I. (1954). *How to lie with statistics*. W. W. Norton & Company: New York.

Knowles, M.S. (1973). The adult learner: A neglected species. Houston, TX: Gulf Publishing Company.

Ladd, E. C. (January 01, 1994). The holocaust poll error: A modern cautionary tale. *Public Perspective, 5,* 5.)

Langrall, C., Nisbet, S., Mooney, E., & Jansem, S. (2011). The role of context expertise when comparing data. *Mathematical Thinking and Learning,13*(1-2), 47-67.

Libman, Z. (July 01, 2008). Integrating real-life data analysis in teaching descriptive statistics: A constructivist approach. *Journal of Statistics Education, 18,* 1.)

Madden, S. R. (2011). Statistically, technologically, and contextually provocative tasks: Supporting teachers' informal inferential reasoning. *Mathematical Thinking and Learning,13*(1-2), 109-131.

McClain, K., & Cobb, P. (February 01, 2001). Supporting students' ability to reason about data. *Educational Studies in Mathematics, 45,* 1-3.

Moll, L. C. (1990). Introduction. In L. C. Moll, *Vygostsky and education: Instructional implications and applications of sociohistorical psychology* (pp. 1-31). New York: Cambridge University Press.

Nussbaum, E. (2002). Scaffolding argumentation in the social studies classroom. *Social Studies*, *93*(3), 79.

Osana, H. P., Leath, E. P., & Thompson, S. E. (2004). Improving evidential argumentation through statistical sampling: evaluating the effects of a classroom intervention for at-risk 7th-graders. *The Journal of Mathematical Behavior,23*(3), 351-370.

Perelman, C. (1982). *The realm of rhetoric*. University of Notre Dame Press: London.

Perelman, C., & Olbrechts-Tyteca, L. (1958). *The new rhetoric: A treatise on argumentation.* University of Notre Dame Press: London.

Pfannkuch, M. (2006). Informal inferential reasoning. Paper presented at the 7th International Conference on Teaching Statistics, July, 2006, Salvador, Brazil.

Pfannkuch, M. (2011). The role of context in developing informal statistical inferential reasoning: A classroom study. *Mathematical Thinking and Learning,13*(1-2), 27-46.

Pfannkuch, M., Forbes, S., Harraway, J., Budgett, S., & Wild, C. (2013). *"Bootstrapping" students' understanding of statistical inference*. online], Research report for Teaching & Learning Initiative, Wellington, New Zealand. Available at http://www.tlri.org.nz/sites/default/files/projects/9295_summary%20report.pdf.

Pfannkuch, M., & Wild, C. (2004). Towards an understanding of statistical thinking. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 17-46). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Pressley, M., Hogan, K., Wharton-McDonald, R., Mistretta, J. and Ettenberger, S. (1996). The challenges of instructional scaffolding: The challenges of instruction that supports student thinking. *Learning Disabilities Research & Practice*, *11*(3), 138-146.

Pugalee, D. (2004). A comparison of verbal and written descriptions of students' problem solving processes. *Educational Studies in Mathematics*, 55(1-3), 27-47.

Roehler, L., & Cantlon, D. (1997). Scaffolding: A powerful tool in social constructivist classrooms. In K. Hogan & M. Pressley (Eds.), *Scaffolding student learning: Instructional approaches and issues* (pp. 6-42). Brookline Books: Cambridge, MA.

Rowland, R.C. (1987). On defining argument. *Philosophy & Rhetoric, 20*(3), 140-159.

Schwarz, B. (2009). Argumentation and learning. In N. Muller Mirza, & A.-N. Perret-Clermont, *Argumentation and education: Theoretical foundations and practices* (pp. 91-126). New York: Springer.

Steffe, L.P., & Thompson, P.W. (2000). Teaching experiment methodology: Underlying principles and essential elements. In A.E. Kelly & R.A. Lesh, Handbook of research design in mathematics and science education (pp. 267-306). Lawrence Erlbaum Associates Publishers: Mahwah, NJ.

Toulmin, S.E. (1958). *The uses of argument*. Cambridge University Press: London.

van de Pol, J.; Volman, M; & Beishuizen, J. (2010). Scaffolding in teacher-student interaction: A decade of research. Educational Psychology Review, *22*, 271-296.

Walker, J., & Sampson, V. (2013). Learning to argue and arguing to learn: argument-driven inquiry as a way to help undergraduate chemistry students learn how to construct arguments and engage in argumentation during a laboratory course. *Journal of Research in Science Teaching*, *50*(5), 561-596.

Watson, J. M. (November 01, 2002). Inferential reasoning and the influence of cognitive conflict. *Educational Studies in Mathematics: An International Journal*, *51*, 3, 225-256.

Wood, D., Bruner, J, & Ross, G. (1976). The role of tutoring in problem solving. Journal of Child Psychology & Psychiatry & Allied Disciplines, *17*(2), 89-100.

Word Press. (n.d.). *Statistical thinking: Psychology, methodology, and technology for overcoming uncertainty*. Retrieved June 5, 2011, from http://statisticalthinking.org/?p=1.

Yackel, E., & Cobb, P. (1996). Sociomathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education*, *27*(4), 458-477.

Zembal-Saul, C., Munford, D., Crawford, B., Friedrichsen, P., & Land, S. (2002). Scaffolding preservice science teachers' evidence-based arguments during an investigation of natural selection. *Research in Science Education*, *32*(4), 437-463.

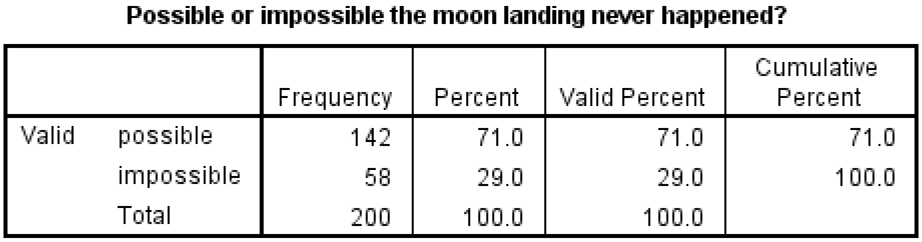
APPENDIX A: INSTRUCTIONAL MATERIALS RELATED TO STATISTICAL ARGUMENTATION

Argument 1 Example

Research question: Do a substantial number of Americans believe the 1969 moon landing was a hoax?

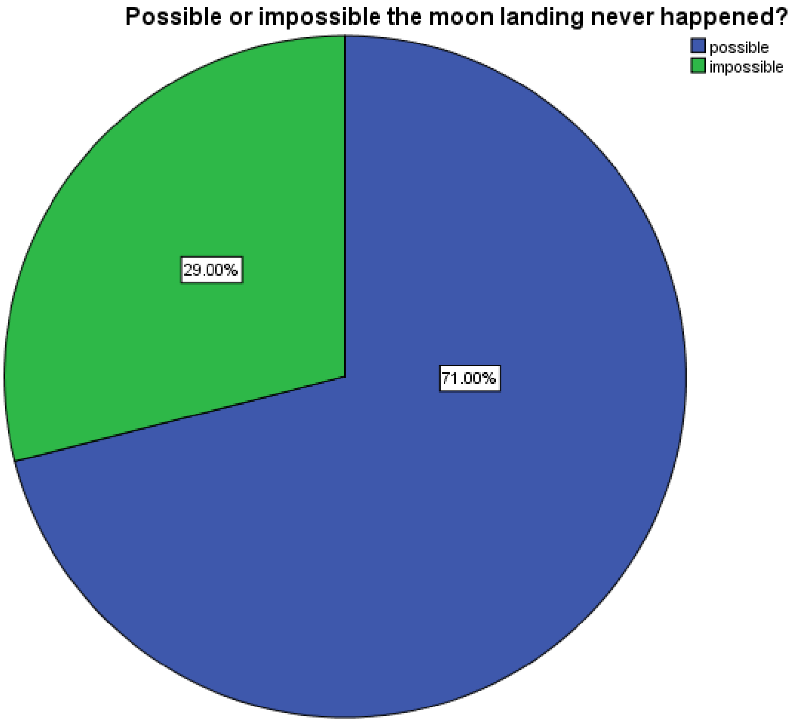
Sampling and data collection procedure: A well-respected polling organization asked a large sample of randomly selected Americans, “Does it seem possible or impossible that the 1969 moon landing in which Neil Armstrong famously said ‘One small step for man, one giant leap for mankind’ never happened?”

Results:



Possible or impossible the moon landing never happened?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	possible	142	71.0	71.0	71.0
	impossible	58	29.0	29.0	100.0
	Total	200	100.0	100.0	



Sample argument: Since the polling organization is well-respected and we are not given further information about sampling techniques, there is no reason to think the sample is biased. It would be beneficial to have more information about sampling, though. While the survey question does appear to address the research question, it is terribly confusing and likely to elicit responses that do not reflect participants' true opinions on the subject. The result of 71% of respondents thinking it is possible the moon landing never happened seems higher than expected, so it is probably more a result of the confusing question than people's true thoughts on the subject. If it is true that 71% of respondents think the moon landing may never have happened, then yes, it is an indication of a substantial number of Americans believing that way. However, more data is required from a survey with a more clearly phrased question before it can be determined what Americans believe about the moon landing.

Elements to include in arguments 1a and 1b

For argument 1, following is a list of things you should consider. A good argument will do most of these. Your argument should be in a well-written paragraph or paragraphs (no bullets!).

Linking to context:

- Assesses whether the research question and data collection procedures make sense for the scenario being discussed.
- States units where appropriate. For the first argument, units are percentages.
- Assesses whether the results are reasonable.

Articulating results:

- Discusses the results of the analysis.

Making inferences:

- Assesses how well the sampling procedures yield a representative sample.
- Discusses how any problems with sampling or data collection limit or otherwise affect conclusions.
- If applicable, makes recommendations to modify the study so research question can be better answered.

Synthesizing:

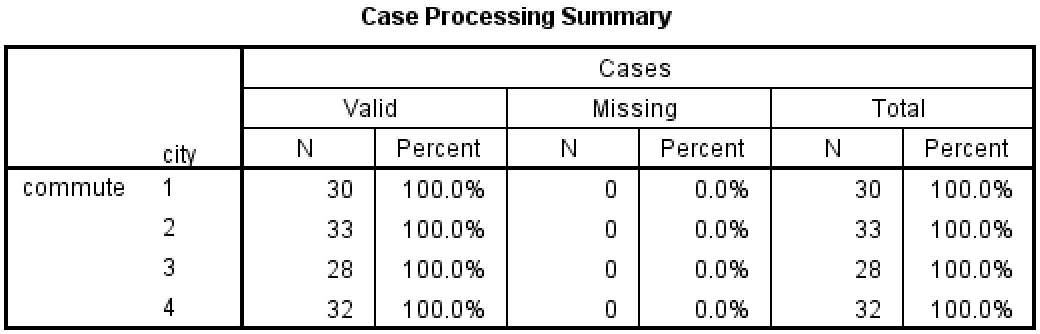
- Synthesizes all of the above elements into a coherent argument that flows.
- Gives a final summary statement that answers the research question.

Argument 2 Example

Research question: How far do people in four cities commute to work?

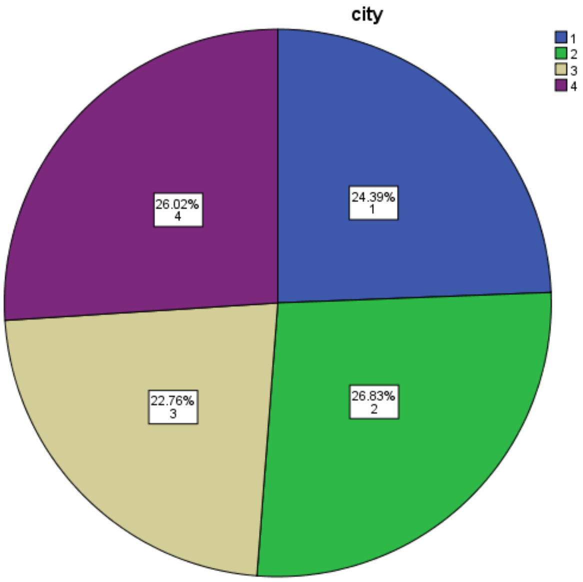
Sampling and data collection procedure: Newspapers in four cities put a question on their websites asking people "How far (in miles) do you drive to work every day one way?" Website visitors were allowed to type in a number to answer the question.

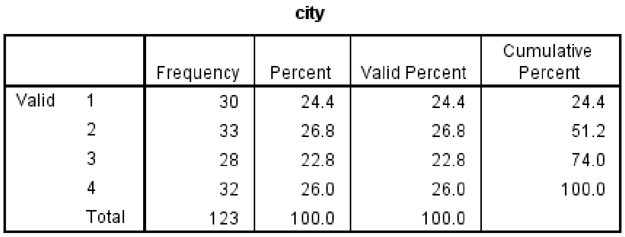
Results:



Case Processing Summary

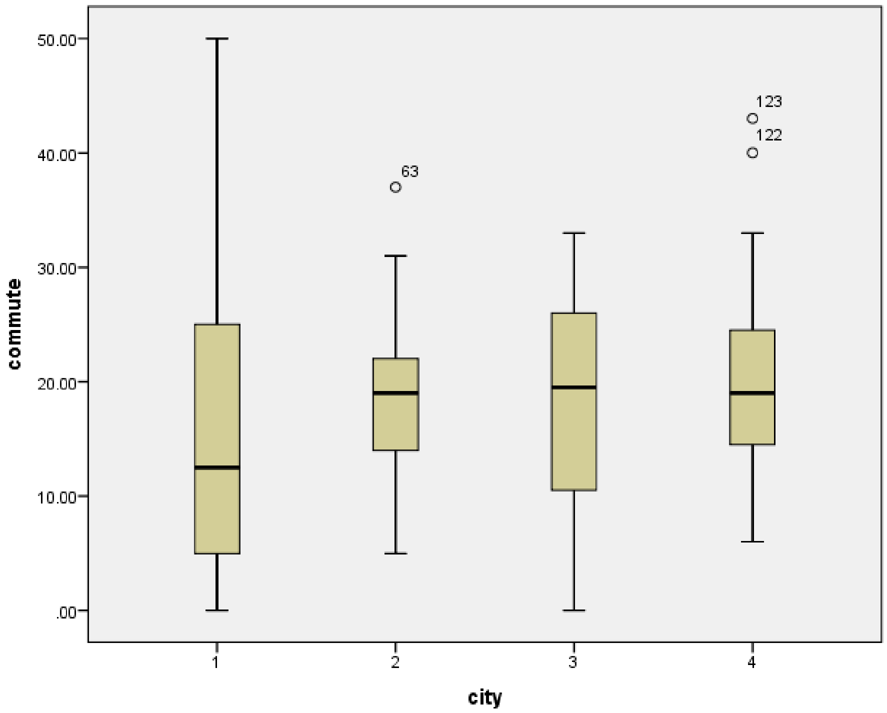
		Cases					
		Valid		Missing		Total	
	city	N	Percent	N	Percent	N	Percent
commute	1	30	100.0%	0	0.0%	30	100.0%
	2	33	100.0%	0	0.0%	33	100.0%
	3	28	100.0%	0	0.0%	28	100.0%
	4	32	100.0%	0	0.0%	32	100.0%

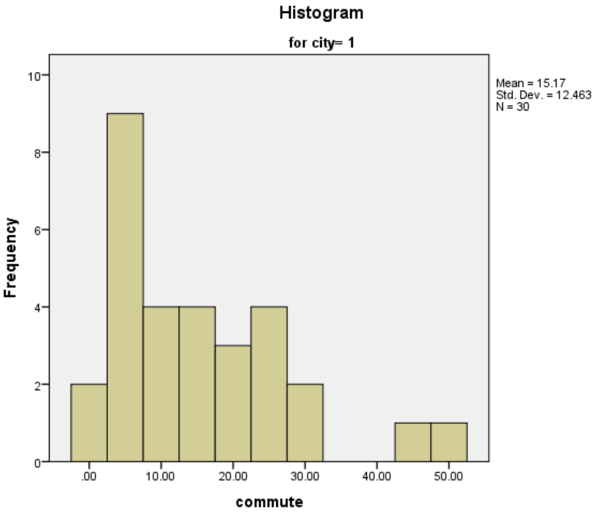


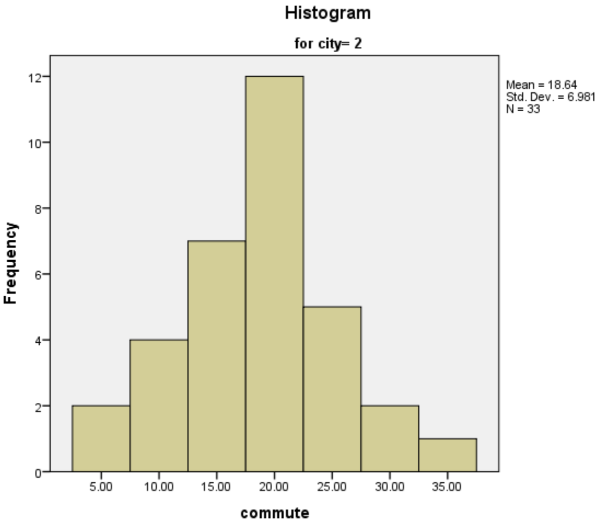


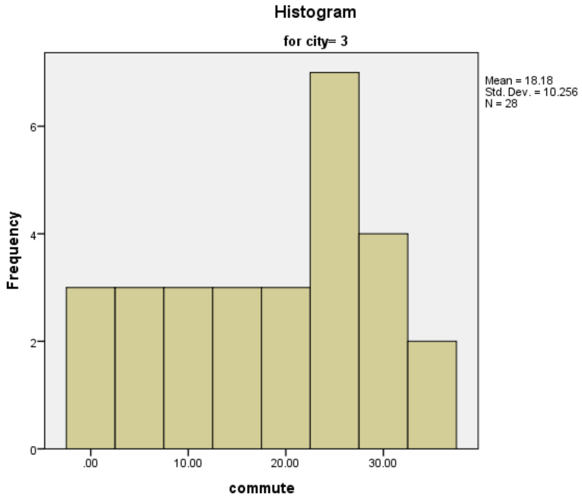
city

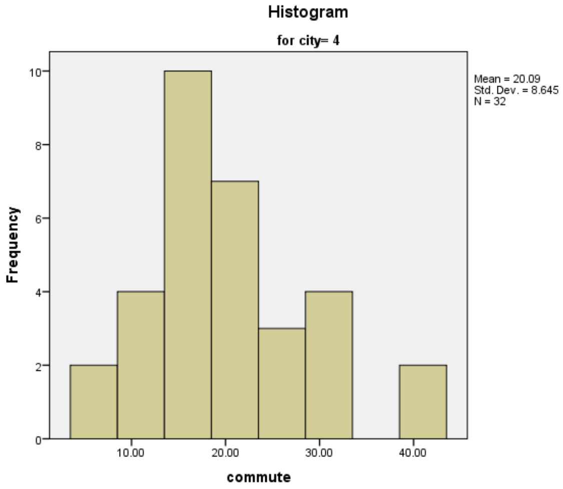
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	30	24.4	24.4	24.4
	2	33	26.8	26.8	51.2
	3	28	22.8	22.8	74.0
	4	32	26.0	26.0	100.0
	Total	123	100.0	100.0	

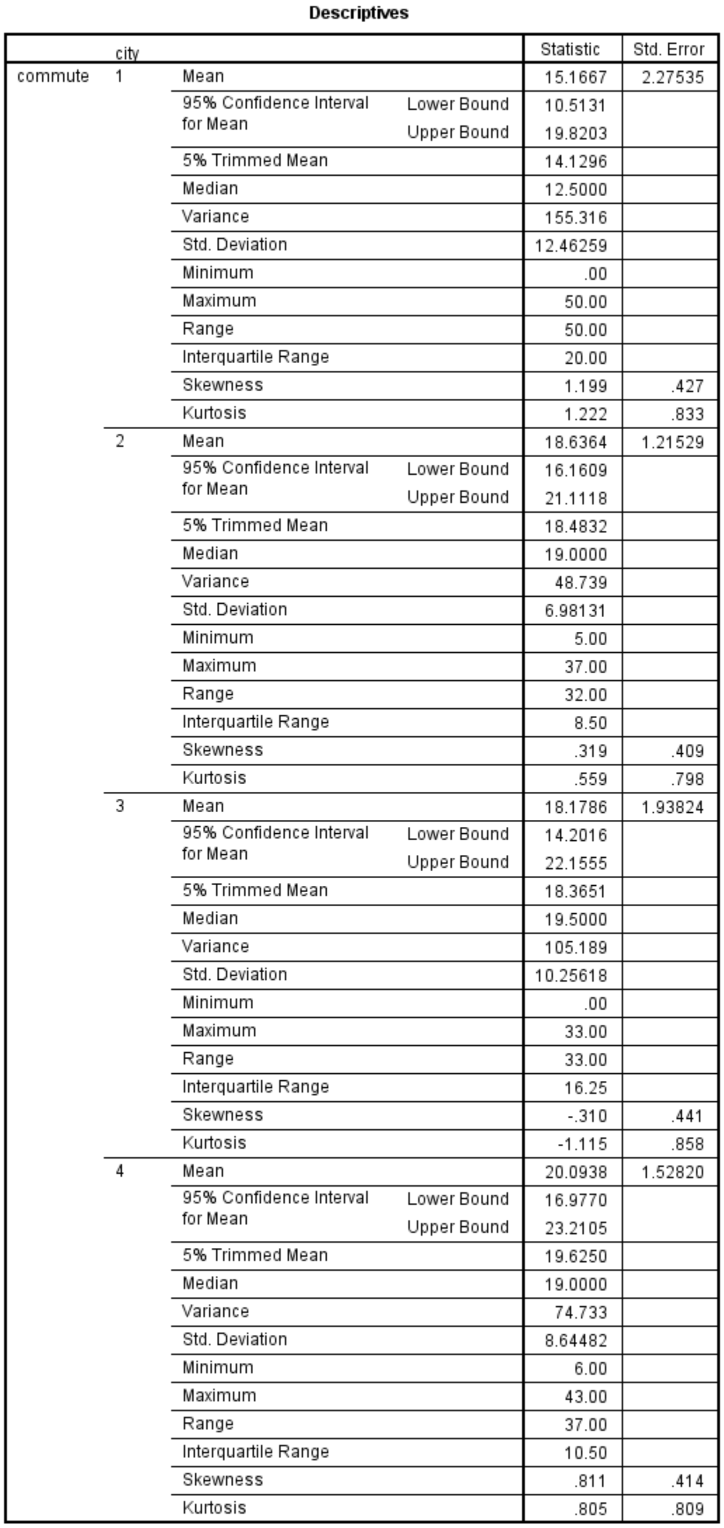

50.00
40.00
30.00
20.00
10.00
.00
commute
63
123
122
1
2
3
4
city


Histogram
for city= 1
Mean = 15.17
Std. Dev. = 12.463
N = 30
Frequency
0
2
4
6
8
10
.00
10.00
20.00
30.00
40.00
50.00
commute


Histogram
for city= 2
Mean = 18.64
Std. Dev. = 6.981
N = 33
Frequency
0
2
4
6
8
10
12
5.00
10.00
15.00
20.00
25.00
30.00
35.00
commute


Histogram
for city= 3
Mean = 18.18
Std. Dev. = 10.256
N = 28
Frequency
0
2
4
6
.00
10.00
20.00
30.00
commute


Histogram
for city= 4
Mean = 20.09
Std. Dev. = 8.645
N = 32
Frequency
0
2
4
6
8
10
10.00
20.00
30.00
40.00
commute



Descriptives

	city			Statistic	Std. Error
commute	1	Mean		15.1667	2.27535
		95% Confidence Interval for Mean	Lower Bound	10.5131	
			Upper Bound	19.8203	
		5% Trimmed Mean		14.1296	
		Median		12.5000	
		Variance		155.316	
		Std. Deviation		12.46259	
		Minimum		.00	
		Maximum		50.00	
		Range		50.00	
		Interquartile Range		20.00	
		Skewness		1.199	.427
		Kurtosis		1.222	.833
	2	Mean		18.6364	1.21529
		95% Confidence Interval for Mean	Lower Bound	16.1609	
			Upper Bound	21.1118	
		5% Trimmed Mean		18.4832	
		Median		19.0000	
		Variance		48.739	
		Std. Deviation		6.98131	
		Minimum		5.00	
		Maximum		37.00	
		Range		32.00	
		Interquartile Range		8.50	
		Skewness		.319	.409
		Kurtosis		.559	.798
	3	Mean		18.1786	1.93824
		95% Confidence Interval for Mean	Lower Bound	14.2016	
			Upper Bound	22.1555	
		5% Trimmed Mean		18.3651	
		Median		19.5000	
		Variance		105.189	
		Std. Deviation		10.25618	
		Minimum		.00	
		Maximum		33.00	
		Range		33.00	
		Interquartile Range		16.25	
		Skewness		-.310	.441
		Kurtosis		-1.115	.858
	4	Mean		20.0938	1.52820
		95% Confidence Interval for Mean	Lower Bound	16.9770	
			Upper Bound	23.2105	
		5% Trimmed Mean		19.6250	
		Median		19.0000	
		Variance		74.733	
		Std. Deviation		8.64482	
		Minimum		6.00	
		Maximum		43.00	
		Range		37.00	
		Interquartile Range		10.50	
		Skewness		.811	.414
		Kurtosis		.805	.809

Sample argument:

The sampling and data collection procedures are problematic. It is a voluntary response sample that relies on visitors to a website who volunteer to answer the question. It is possible that these participants are different from people who did not participate in the survey. Further, there is no guarantee that participants even lived in the cities being studied.

The four cities in the dataset are represented roughly equally, as seen in the pie chart and accompanying frequency table. The boxplots show approximately normal distributions for each of the four cities, with one outlier in city 2 on the high side and two outliers in city 4 on the high side. From the boxplots, all four cities appear to be similar in center, but city 1 seems to have a higher spread than the other three. Because the boxplots and histograms do not show any strong skewness for any of the cities, the means are the best measures of center. The means of cities 2 and 3 are almost identical at 18.6 miles and 18.2 miles. City 1 has a smaller mean commute distance of 15.2 miles and city 4 has a larger mean of 20.1 miles. The standard deviations are 12.5 miles for city 1 (CV=82%), 7.0 miles for city 2 (CV=37%), 10.3 miles for city 3 (CV=56%), and 8.6 miles for city 4 (CV=43%). These values indicate dissimilarity among the four cities, with city 1 having the highest variability.

From the data collected, it appears that people in the four cities commute anywhere from 0 to 50 miles to work one way, with the average being about 18 miles. City 1 has the shortest commute on average, 15 miles, and the largest variability. Cities 2 and 3 have average commute distances of approximately 18.5 miles, while city 4 has an average commute of about 20 miles. These values are only for the sample, however. Inferences to the populations of these cities must be made with extreme caution, since the sampling procedures were likely to yield a biased sample. Random samples might provide better information about the populations in these four cities.

Elements to include in arguments 2a and 2b

For argument 2, following is a list of things you should consider. A good argument will do most of these. Your argument should be in a well-written paragraph or paragraphs (no bullets!).

Linking to context:

- Assesses whether the research question and data collection procedures make sense for the scenario being discussed.
- States units where appropriate. For example, if the units are mph and the mean is 42, state the mean as 42 mph.
- Assesses whether the results are reasonable.

Articulating results:

- Discusses appropriate measures of center of the data. Compares and contrasts across groups.
- Discusses appropriate measures of spread of the data. Compares and contrasts across groups.
- Discusses the distribution of the data, including distribution shape and outliers. Compares and contrasts across groups.

Making inferences:

- Assesses how well the sampling procedures yield a representative sample.
- Discusses how any problems with sampling or data collection limit or otherwise affect conclusions.
- If applicable, makes recommendations to modify the study so research question can be better answered.

Synthesizing:

- Synthesizes all of the above elements into a coherent argument that flows.
- Gives a final summary statement that answers the research question.

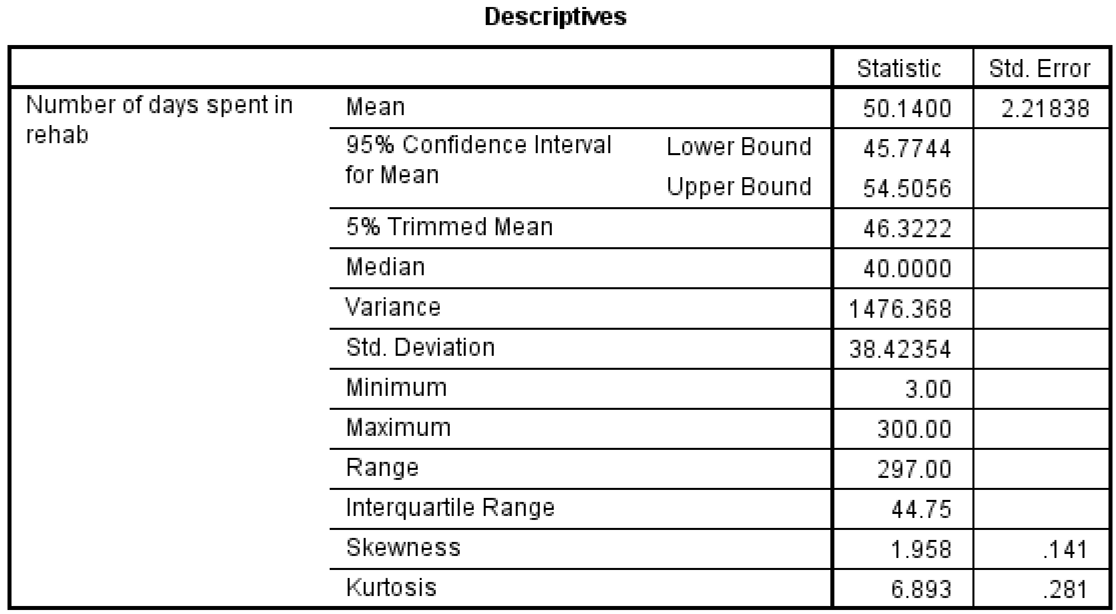
Statistical Argument #3 example

Suppose a study is conducted and you are provided with the following information.

Research question: A state legislature is trying to make a law requiring insurance companies to cover more days of in-patient rehab for substance abuse. One lawmaker has proposed insurance companies be required to cover at least 55 days. The legislative committee in charge of vetting the bill wants to make sure 55 is more than the mean number of days in rehab needed for full recovery. According to this standard, is 55 days a large enough number?

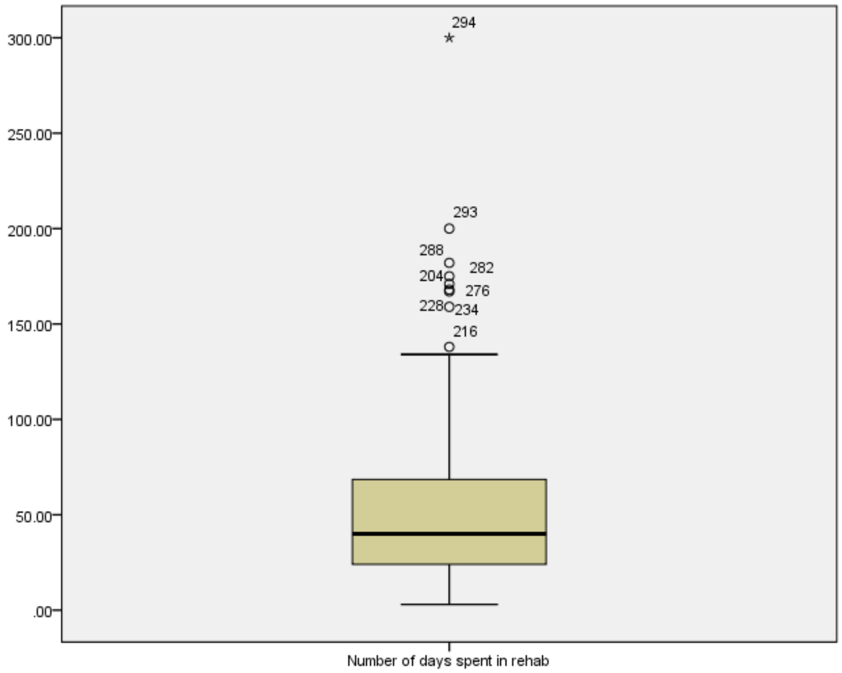
Sampling and data collection procedure: Six facilities in the state were randomly selected using a random number generator. Administrators at each of the six facilities were sent instructions for how to select a sample (again, using a random number generator) of 50 recent patients. They recorded the number of days in rehab for each of the 50 patients in the survey.

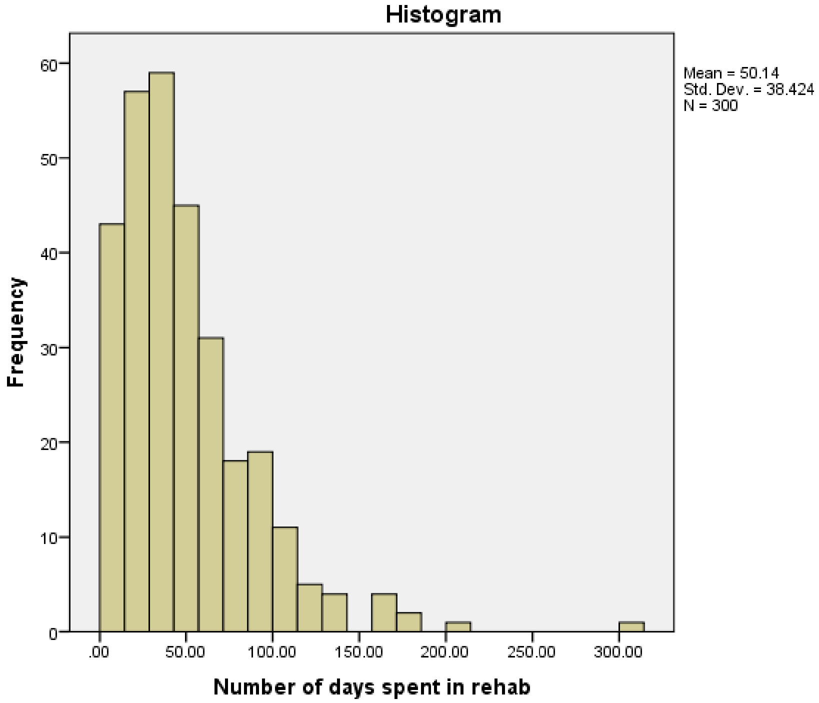
Results:



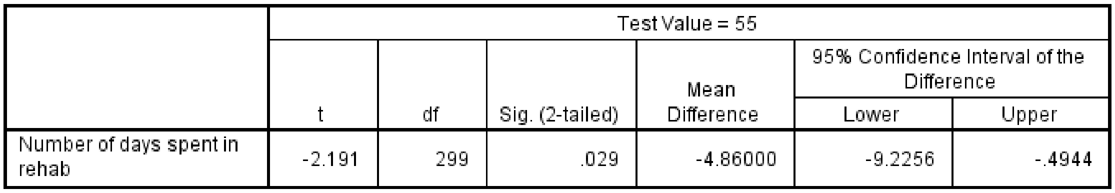
Descriptives

			Statistic	Std. Error
Number of days spent in rehab	Mean		50.1400	2.21838
	95% Confidence Interval for Mean	Lower Bound	45.7744	
		Upper Bound	54.5056	
	5% Trimmed Mean		46.3222	
	Median		40.0000	
	Variance		1476.368	
	Std. Deviation		38.42354	
	Minimum		3.00	
	Maximum		300.00	
	Range		297.00	
	Interquartile Range		44.75	
	Skewness		1.958	.141
	Kurtosis		6.893	.281





One-Sample Test



	Test Value = 55					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Number of days spent in rehab	-2.191	299	.029	-4.86000	-9.2256	-.4944

Sample argument:

Sampling and data collection procedures appear mostly sound. Random selection of facilities and patients virtually guarantees a representative sample. There is potential for motivation to lie, since the facilities are reporting their own numbers. Additionally, the results are being used to require insurance companies to cover more days of rehab, which helps the patients at the facility and benefits the facility financially. However, most of the observations are low—half are below 40 days—which makes the data seem like it is not inflated. This study does have a fundamental problem, though, which is that many of the patients who attend rehab have insurance and will probably only stay as long as their insurance will cover. Thus it is not possible to know how long these patients needed to stay or would have stayed.

The data ranges from 3 days to 300 days, a very large span of values. The data is strongly skewed to the right, as shown by both the histogram and the boxplot. There are nine outliers ranging from approximately 140 days to 300 days. While 300 days is a long time to spend in rehab and it may be a data entry error, it may also be a legitimate data point. The mean of the data is 50.14 days. Because the data is skewed, the median of 40 days is a better measure of center. At 38.42 days (CV=76.63%), the standard deviation is large compared to the mean, which further indicates a large spread.

Though the data does not follow a normal distribution, the sample size of 300 is large enough that it is still valid to conduct the t-test for the mean. The hypothesis test of H_0: μ=55 vs. H_1: μ<55 has a p-value of .0145. At the .05 level of significance, we would reject H_0, meaning there is sufficient evidence to conclude the population mean is less than 55 days. At .1265, the effect size (Cohen's *d*) is small, indicating that even though we were able to conclude the mean number of days of all patients is less than 55 days, the difference is minor.

As far as this study is concerned, 55 days is an appropriate length of stay in rehab to require insurance companies to cover. However, because of the limitations of the study, it is impossible to determine what the mean number of days would have been if patients had not been limited by insurance. Before making a final decision, the state legislature should do more research to determine not only how long patients currently spend in rehab but also how long they need to spend in rehab.

Elements to include in arguments 3a and 3b

For argument 3, following is a list of things you should consider. A good argument will do most of these. Your argument should be in a well-written paragraph or paragraphs (no bullets!).

Linking to context:

- Assesses whether the research question and data collection procedures make sense for the scenario being discussed.
- States units where appropriate.
- Assesses whether the results are surprising.

Articulating results:

- Discusses appropriate measures of center and spread of the data. Compares across groups if the data has multiple groups.
- Discusses the distribution of the data, including distribution shape and outliers.
- Carries out hypothesis tests, and discusses associated effect sizes.
- Notes whether assumptions are met for hypothesis tests.

Making inferences:

- Assesses how well the sampling procedures yield a representative sample.
- Makes conclusions about the population based on the information in the sample.
- Discusses how any problems with sampling or data collection limit or otherwise affect conclusions.
- Makes recommendations to modify the study so research question can be better answered.

Synthesizing:

- Synthesizes all of the above elements into a coherent argument that flows.
- Gives a final summary statement that answers the research question.

Notes for t-tests for statistical reasoning assignments 3a and 3b

Step 1: Evaluate assumptions

For the t-test, we need either a large sample size (at least 30) or sampling from a normal distribution (based on the histogram).

Step 2: Conduct the test

Possible hypotheses are H_0: μ= vs. H_1: μ<(fill in the value), or H_0: μ= vs. H_1: μ>(fill in the value).

If the sample mean is less than the hypothesized value, the alternative hypothesis will be μ< (hypothesized value).

If the sample mean is greater than the hypothesized value, the alternative hypothesis will be μ> (hypothesized value).

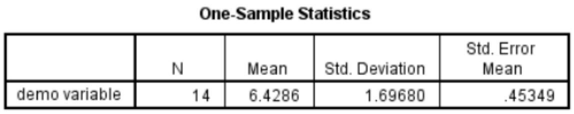
The p-value for the test is the “Sig. (2-tailed) value” on the SPSS output divided by 2.

Use $\alpha = .05$.

If you reject H_0, the result is considered “statistically significant”.

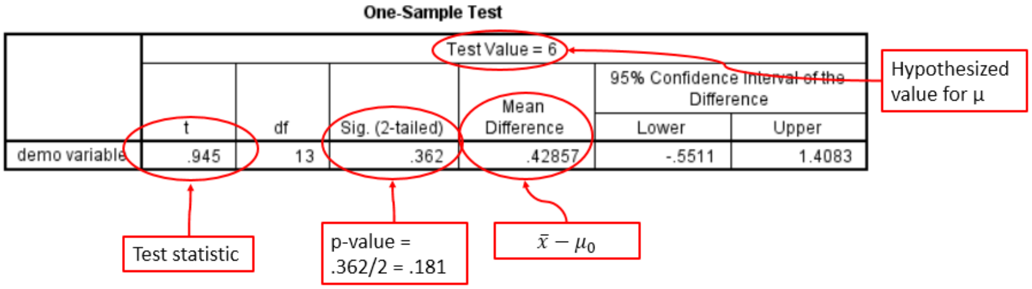
Step 3: Calculate and interpret the effect size.

The effect size, called Cohen’s *d*, is calculated as $d = \frac{|\bar{x}-\mu_0|}{s}$. In the SPSS output, $\bar{x} - \mu_0$ is found under “Mean Difference”. Rough guidelines are: .2 is small, .5 is medium, and .8 is large.



One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
demo variable	14	6.4286	1.69680	.45349



One-Sample Test

	Test Value = 6					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
demo variable	.945	13	.362	.42857	-.5511	1.4083

Hypothesized value for µ
Test statistic
p-value = .362/2 = .181
$\bar{x} - \mu_0$

Since the sample mean is greater than the hypothesized value, this is a test of H_0: µ=6 vs. H_0: µ>6.
The p-value is .181.
The effect size is | .42857 | / 1.69680 = .2526.

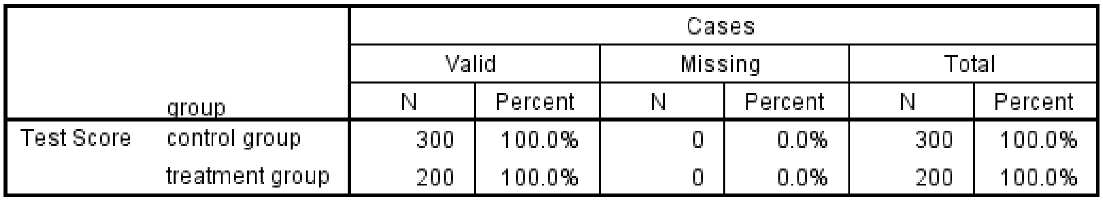
Statistical Argument 4 example

Research question: Does teacher perception of student aptitude impact the academic achievement of students?

Sampling and data collection procedure: In a group of 500 rising third-graders, 200 were randomly selected via random number generator to participate in an experiment. At the beginning of the school year, teachers were told that these 200 students (the treatment group) were exceptionally intelligent, regardless of the students' prior grades or test scores. At the end of the academic year, their scores on an achievement test were measured and compared to the scores of the other students (the control group). Scores for this achievement test are measured on a scale of 1 to 500.

Results:

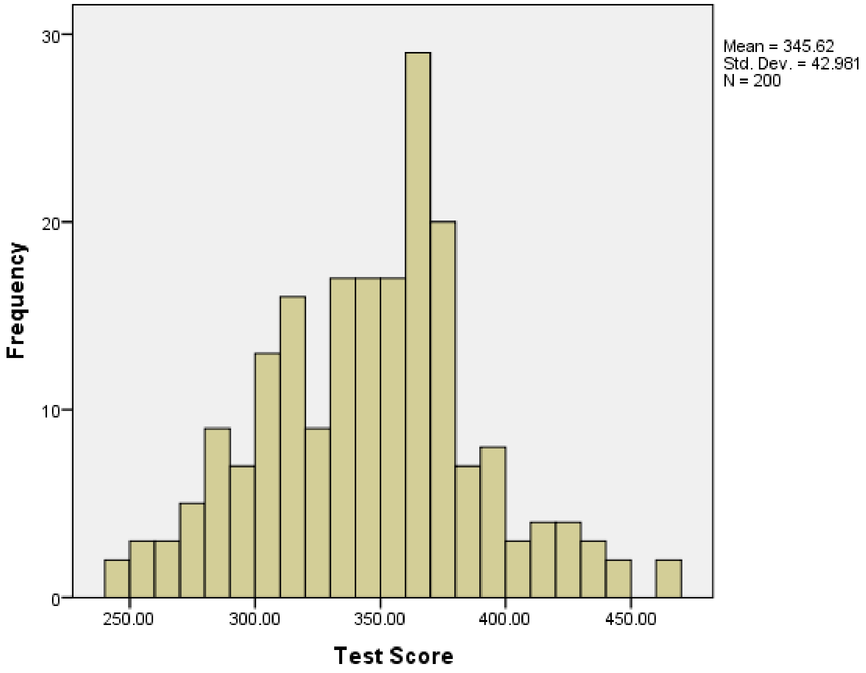
Case Processing Summary

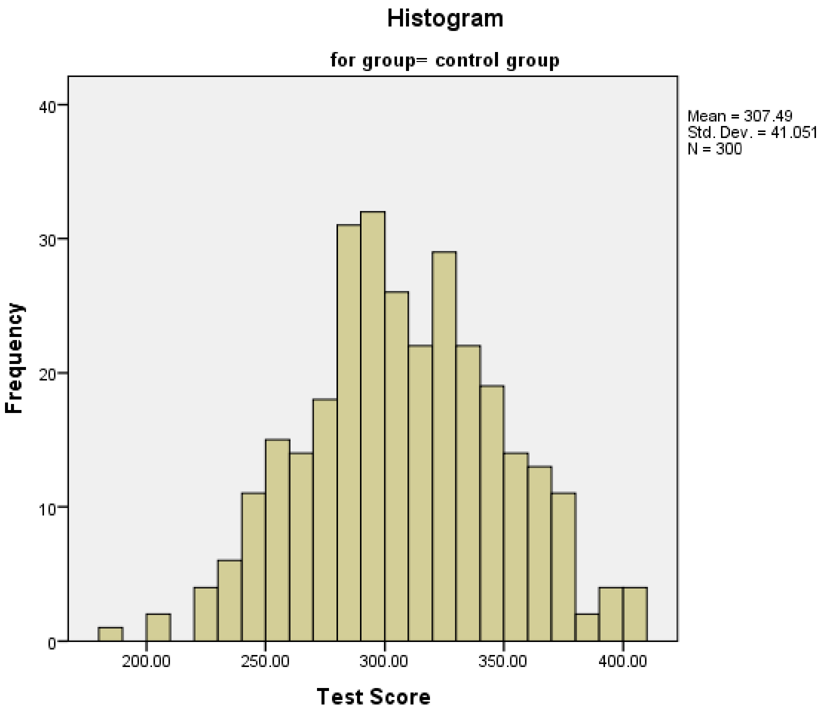


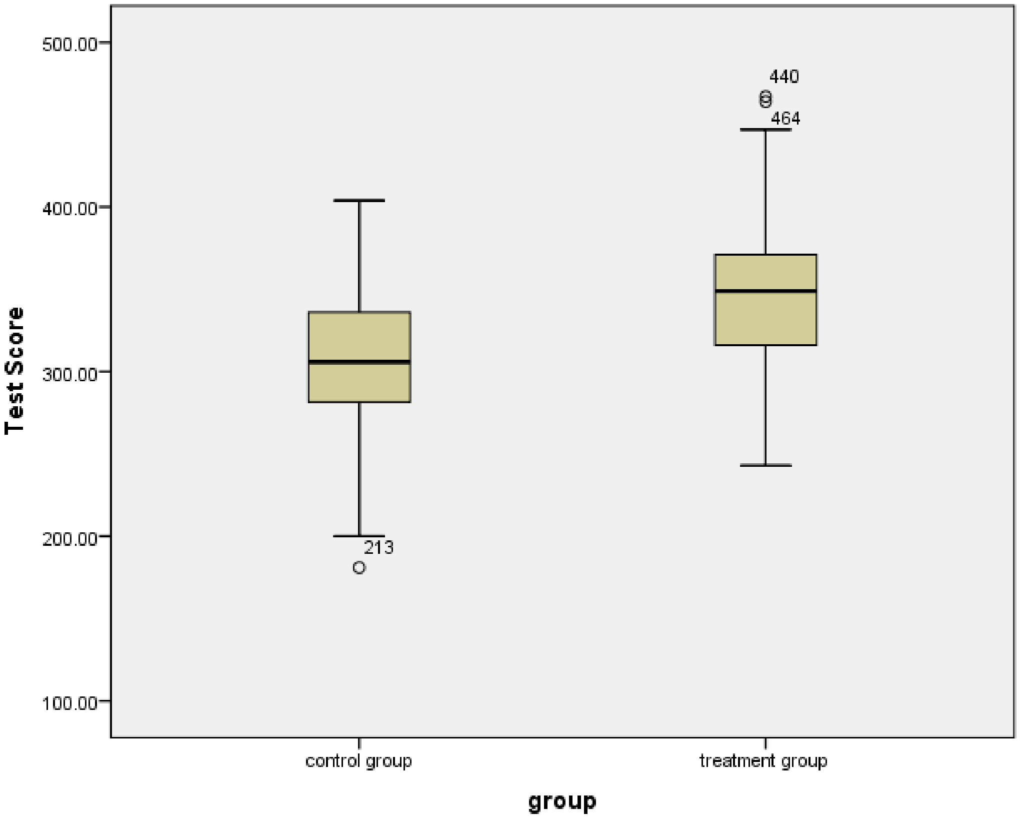
		Cases					
		Valid		Missing		Total	
	group	N	Percent	N	Percent	N	Percent
Test Score	control group	300	100.0%	0	0.0%	300	100.0%
	treatment group	200	100.0%	0	0.0%	200	100.0%

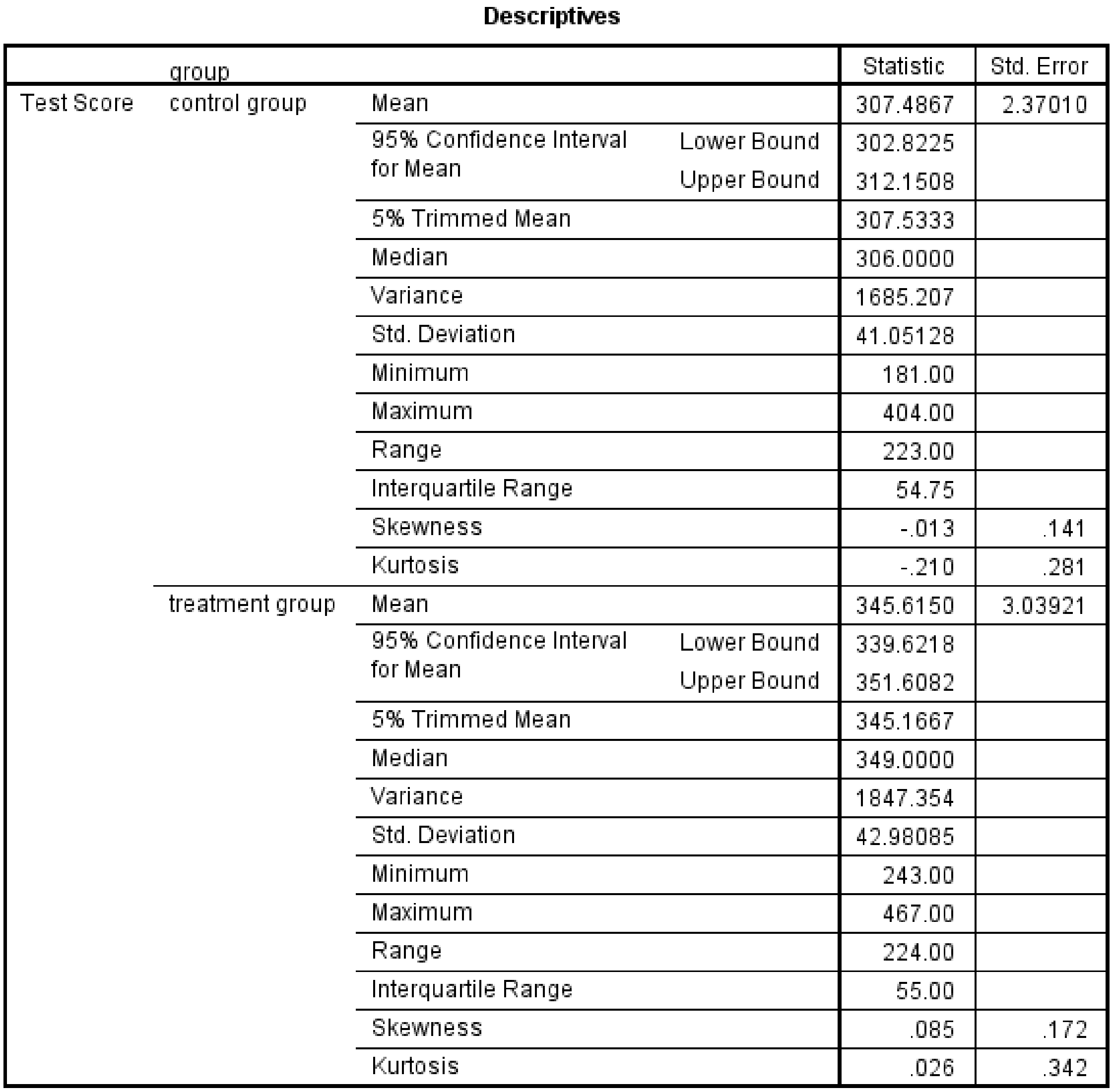
Histogram

for group= treatment group



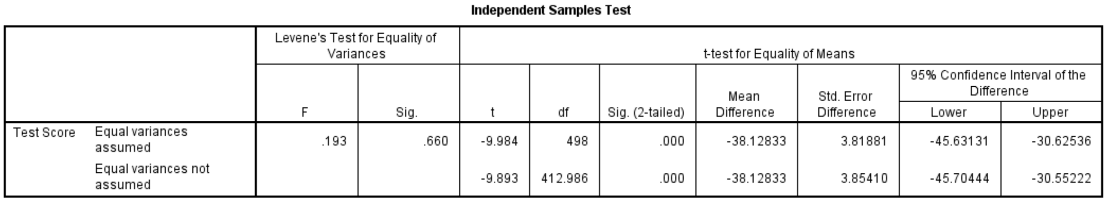

Histogram
for group= control group
Mean = 307.49
Std. Dev. = 41.051
N = 300
Frequency
0
10
20
30
40
200.00
250.00
300.00
350.00
400.00
Test Score


Test Score
500.00
400.00
300.00
200.00
100.00
440
464
213
control group
treatment group
group



Descriptives

	group			Statistic	Std. Error
Test Score	control group	Mean		307.4867	2.37010
		95% Confidence Interval for Mean	Lower Bound	302.8225	
			Upper Bound	312.1508	
		5% Trimmed Mean		307.5333	
		Median		306.0000	
		Variance		1685.207	
		Std. Deviation		41.05128	
		Minimum		181.00	
		Maximum		404.00	
		Range		223.00	
		Interquartile Range		54.75	
		Skewness		-.013	.141
		Kurtosis		-.210	.281
	treatment group	Mean		345.6150	3.03921
		95% Confidence Interval for Mean	Lower Bound	339.6218	
			Upper Bound	351.6082	
		5% Trimmed Mean		345.1667	
		Median		349.0000	
		Variance		1847.354	
		Std. Deviation		42.98085	
		Minimum		243.00	
		Maximum		467.00	
		Range		224.00	
		Interquartile Range		55.00	
		Skewness		.085	.172
		Kurtosis		.026	.342



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Test Score	Equal variances assumed	.193	.660	-9.984	498	.000	-38.12833	3.81881	-45.63131	-30.62536
	Equal variances not assumed			-9.893	412.986	.000	-38.12833	3.85410	-45.70444	-30.55222

Sample argument:

Data collection procedures appear to be appropriate. Random assignment means there should be no difference between the control and treatment groups at the outset of the study. One possible concern is that academic achievement is defined as the score on a single standardized test. Ideally, the researchers would use more than one measure of academic achievement.

From the histograms it appears the test scores of both the treatment and control groups follow approximately a normal distribution. There are three outliers: one in the control group with a low score and two in the treatment group with high scores. None are extreme outliers and all are within the possible range of scores. The mean scores for the treatment group and the control group are 345.6 and 307.5, respectively. The standard deviations are 41.1 and 43.0, which yield coefficients of variation of 11.9% for the control group and 14.0% for the treatment group, indication that the treatment group has more spread than the control group.

Because both sample sizes exceed 30 and both samples have approximately normal distributions, the assumption to conduct the two-sample t test has been met. The two-sample t test of H_0: $\mu_1=\mu_2$ vs. H_1: $\mu_1<\mu_2$, where group 1 is the control group and group 2 is the treatment group, has a p-value that rounds to zero. Using a .05 level of significance, then, we would reject H_0 and conclude that students in the treatment group have higher scores on average than students in the control group. Cohen's *d* was calculated to be .8871, which indicates that the difference is large. Based on this information, when teachers are given information that students are exceptionally intelligent, the students on average perform much better on an end-of-grade test. Because this was an experiment, we can conclude the information given to the teachers was the cause of the higher mean score. Though it would help to see varied measures of academic achievement, this study points to teacher expectations playing a powerful role in student success.

Elements to include in arguments 4a and 4b

For argument 4, following is a list of things you should consider. A good argument will do most of these. Your argument should be in a well-written paragraph or paragraphs (no bullets!).

Linking to context:

- Assesses whether the research question and data collection procedures make sense for the scenario being discussed.
- States units where appropriate. For example, if the units are mph and the mean is 42, state the mean as 42 mph.
- Assesses whether the results are reasonable.

Articulating results:

- Discusses appropriate measures of center of the data. Compares centers of the two groups.
- Discusses appropriate measures of spread of the data. Compares spreads of the two groups.
- Discusses the distribution of each group, including distribution shape and outliers.
- Determines whether assumptions have been met to conduct the hypothesis test.
- Conducts a hypothesis test, including statement of hypotheses, p-value, and conclusion in the words of the problem. The p-value should be based on the SPSS output, not the calculator.
- Calculates and interprets the effect size.

Making inferences:

- Assesses how well the sampling procedures yield a representative sample.
- Discusses how any problems with sampling or data collection limit or otherwise affect conclusions.
- If applicable, makes recommendations to modify the study so research question can be better answered.

Synthesizing:

- Synthesizes all of the above elements into a coherent argument that flows.
- Gives a final summary statement that answers the research question.

Notes for t-tests for statistical reasoning assignments 4a and 4b

Step 1: Evaluate assumptions

For the t-test, we need either for both groups to have a large sample size (at least 30) or for both groups to be sampled from normal distributions (based on the histograms).

Step 2: Conduct the test

Possible hypotheses are H_0: $\mu_1 = \mu_2$ vs. H_1: $\mu_1 < \mu_2$ or H_1: $\mu_1 > \mu_2$.

If the mean for group 1 is less than the mean for group 2, the alternative hypothesis will be $\mu_1 < \mu_2$. If the mean for group 1 is greater than the mean for group 2, the alternative hypothesis will be $\mu_1 > \mu_2$.

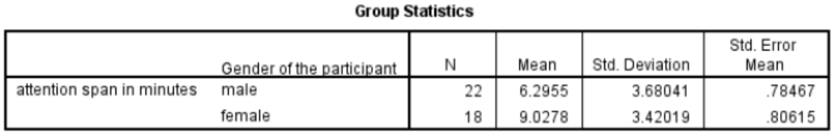
The p-value for the test is the "Sig. (2-tailed) value" on the SPSS output divided by 2. Use the "equal variances not assumed" option.

Use $\alpha = .05$.

If you reject H_0, the result is considered "statistically significant".

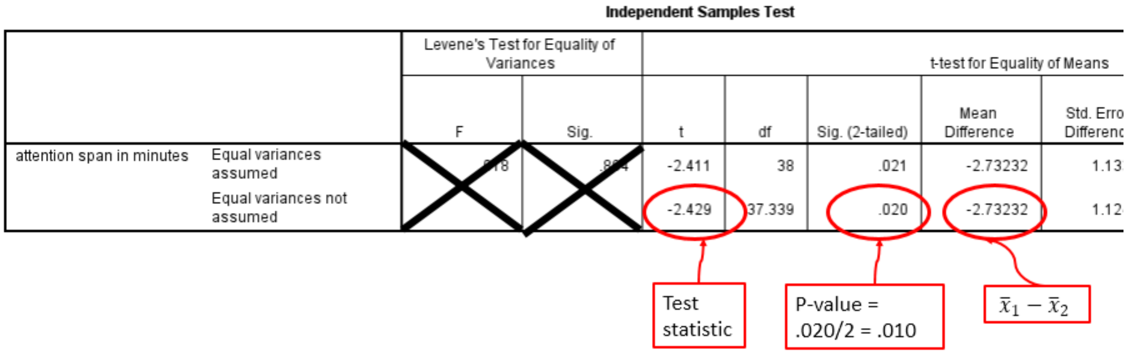
Step 3: Calculate and interpret the effect size.

The effect size for the two-sample t test, still called Cohen's *d*, is calculated as $d = \frac{|\bar{x}_1 - \bar{x}_2|}{s}$, where s is the larger of the two group standard deviations. In the SPSS output, $\bar{x}_1 - \bar{x}_2$ is found under "Mean Difference". Rough guidelines are: .2 is small, .5 is medium, and .8 is large.



Group Statistics

	Gender of the participant	N	Mean	Std. Deviation	Std. Error Mean
attention span in minutes	male	22	6.2955	3.68041	.78467
	female	18	9.0278	3.42019	.80615



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Erro Differenc
attention span in minutes	Equal variances assumed	~~.018~~	~~.894~~	-2.411	38	.021	-2.73232	1.13
	Equal variances not assumed			-2.429	37.339	.020	-2.73232	1.12

Test statistic

P-value = .020/2 = .010

$\bar{x}_1 - \bar{x}_2$

Since group 1 (males) has a lower mean than group 2 (females), the hypotheses are H$_0$: $\mu_1 = \mu_2$ vs. H$_1$: $\mu_1 < \mu_2$.

p-value = .020/2 = .010. Effect size is $d = \frac{|\bar{x}_1 - \bar{x}_2|}{s} = \frac{|-2.73232|}{3.42019}$ = .7989

APPENDIX B: TASKS RELATED TO STATISTICAL ARGUMENTATION

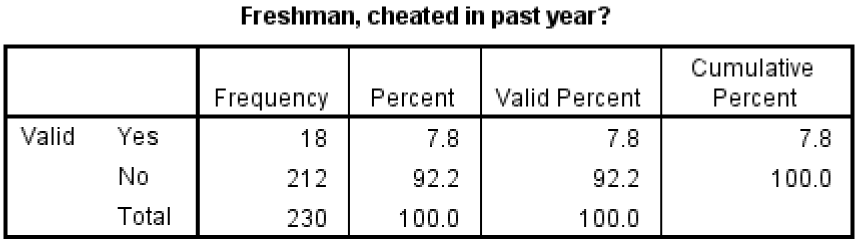
Statistical Reasoning Assignment #1a

Suppose a study is conducted and you are provided with the following information.

Research question: Does academic cheating among college students become more prevalent over the course of their academic careers?

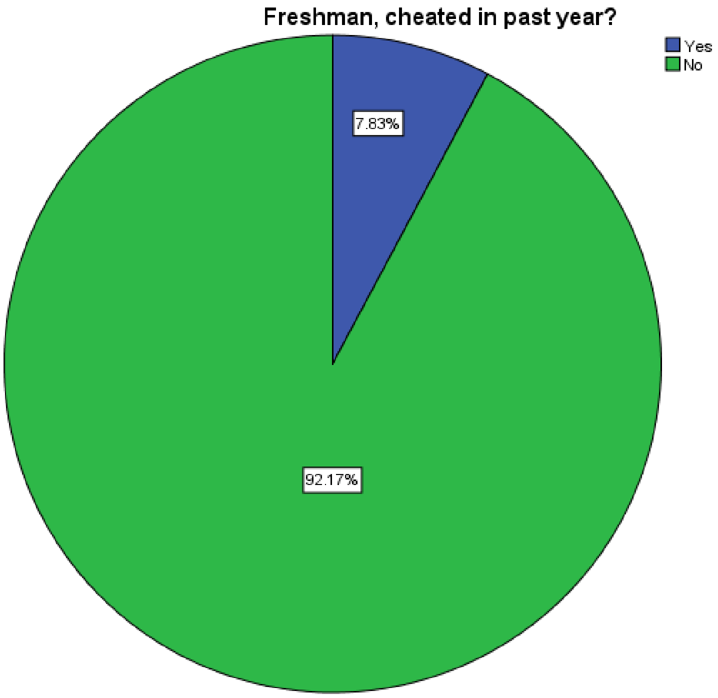
Sampling and data collection procedure: At a small elite college known for a strict honor code, all 250 freshmen were selected to be a sample of all college students in the United States. During freshman orientation, academic advisors asked students, "Academic cheating is defined as gaining or helping another gain an unfair advantage through dishonesty. It may include, for example, plagiarism, giving or receiving unauthorized help on assignments, or forging a grade. In the past year, have you cheated academically?". Three years later, the same group of 250 students was asked the same question their senior year by their academic advisors.

Results: (On the tables below, there are not 250 responses to each question. That's because some of the 250 students refused to answer.)



Freshman, cheated in past year?

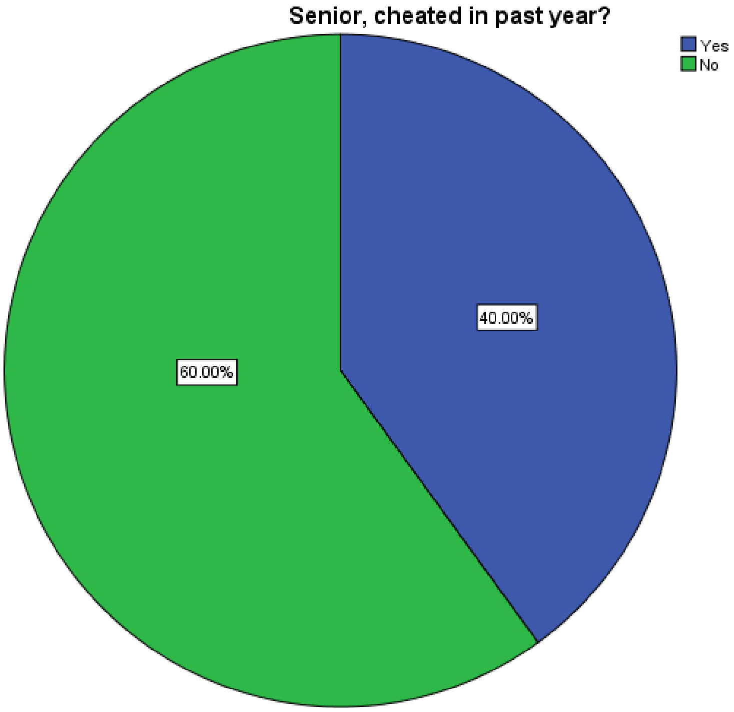
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	18	7.8	7.8	7.8
	No	212	92.2	92.2	100.0
	Total	230	100.0	100.0	



Freshman, cheated in past year?

Senior, cheated in past year?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	14	40.0	40.0	40.0
	No	21	60.0	60.0	100.0
	Total	35	100.0	100.0	



Based on the information provided, work with your assigned group to write an argument that answers the research question.

Expectations:

1. You may discuss this assignment with your assigned group and your instructor only. Any discussion, regardless of how minor, with classmates, tutors, or anyone else will be considered an honor code violation.
2. All group members should be involved in every phase of the work, and all should approve of the final product before it is submitted.
3. Submit your group’s typed argument—just one per group with all names on it—to MyCourses in a .doc, .docx, or .pdf format. If you don’t trust a group member to submit the assignment on time, you may upload more than one per group, as long as group members turn in the exact same work.
4. Please alert me as soon as possible if any member of your group is not participating in the work.

Statistical Reasoning Assignment #1b

Suppose a study is conducted and you are provided with the following information.

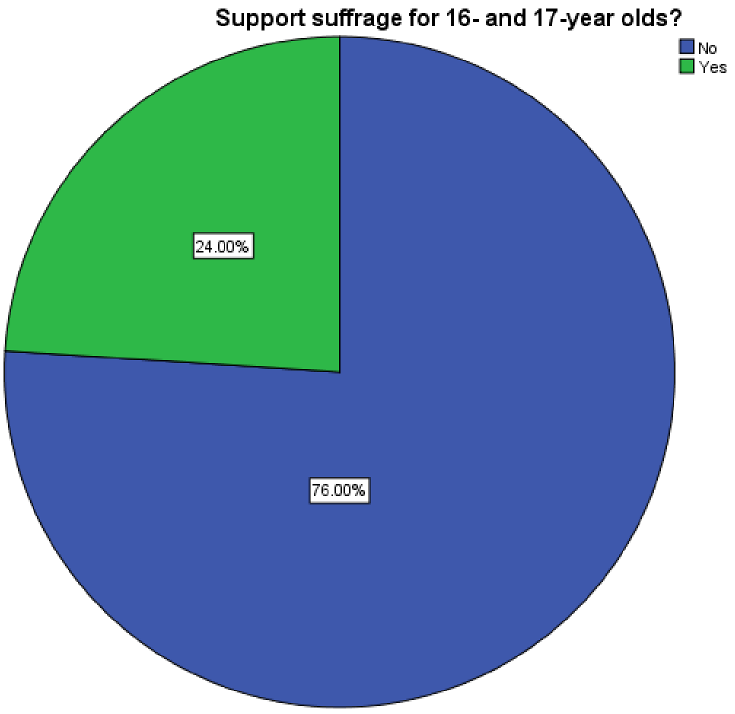
Research question: Do teens pay attention to local and national politics?

Sampling and data collection procedure: Participants were recruited from a popular arcade at a local mall. Fifty teens were asked the question, “Would you support suffrage for 16- and 17-year olds?”

Results:

Support suffrage for 16- and 17-year olds?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	38	76.0	76.0	76.0
	Yes	12	24.0	24.0	100.0
	Total	50	100.0	100.0	



Based on the information provided, work independently to construct an argument that answers the research question.

Expectations:

1. You may discuss this assignment with your instructor only. Any discussion, regardless of how minor, with classmates, tutors, or anyone else will be considered an honor code violation.
2. Submit your typed argument—just one per group with all names on it—to MyCourses in a .doc, .docx, or .pdf format.

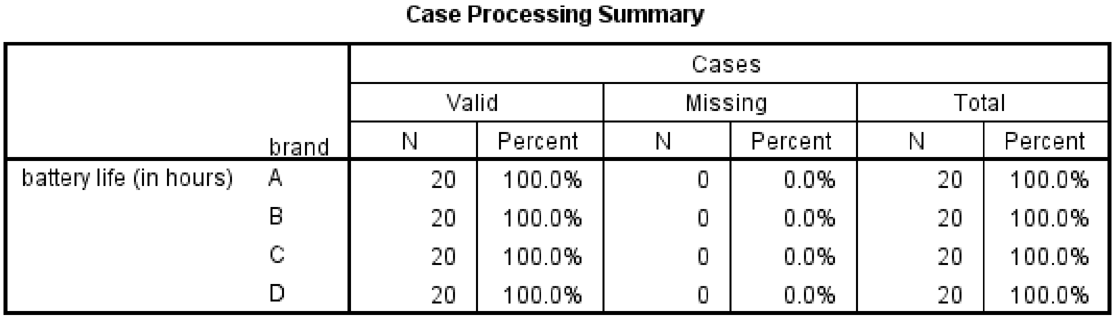
Statistical Reasoning Assignment #2a

Suppose a study is conducted and you are provided with the following information.

Research question: How long do four brands of AAA alkaline batteries work before losing power?

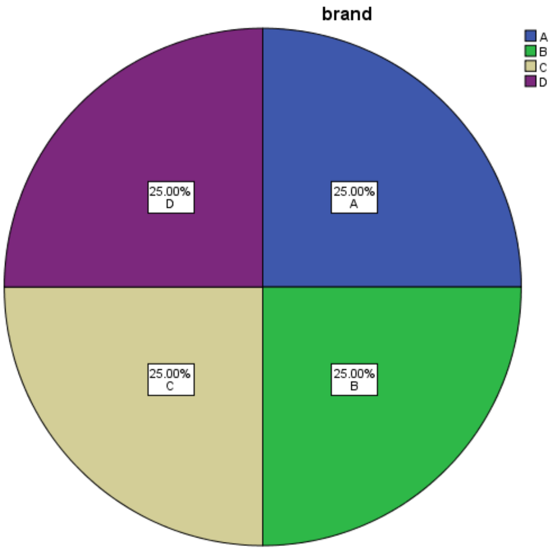
Sampling and data collection procedure: Twenty AAA batteries of each brand are purchased, all from the same store at the same time. The batteries were tested in several different children's toys. For each trial, the amount of time until the battery in the toy stopped functioning was noted.

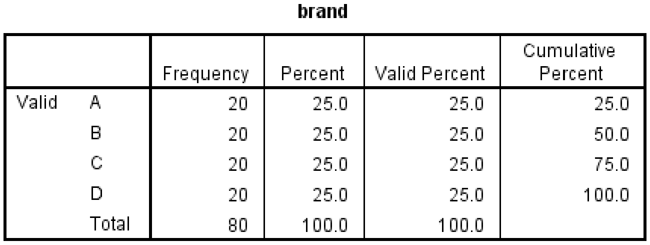
Results:



Case Processing Summary

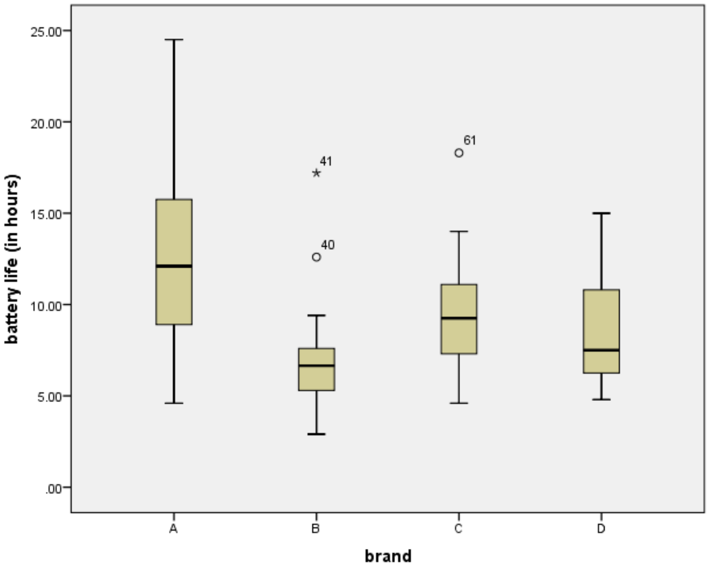
		Cases					
		Valid		Missing		Total	
	brand	N	Percent	N	Percent	N	Percent
battery life (in hours)	A	20	100.0%	0	0.0%	20	100.0%
	B	20	100.0%	0	0.0%	20	100.0%
	C	20	100.0%	0	0.0%	20	100.0%
	D	20	100.0%	0	0.0%	20	100.0%

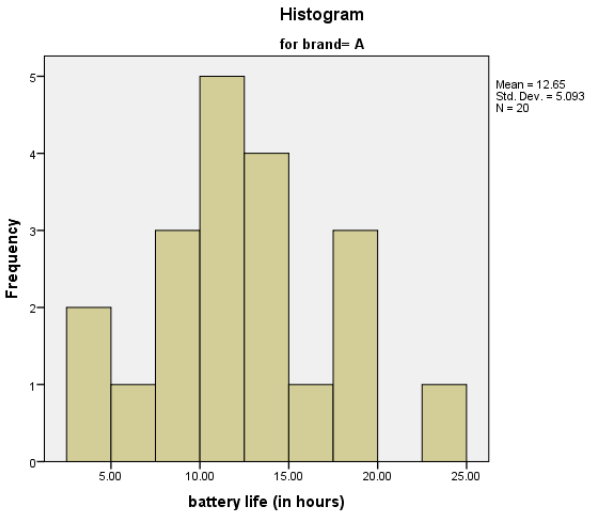


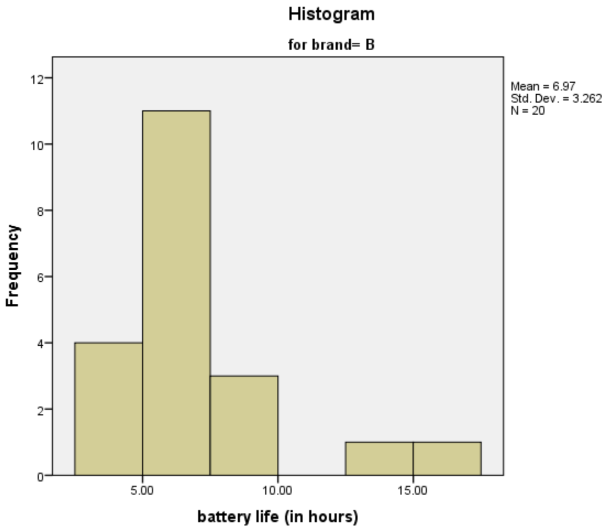


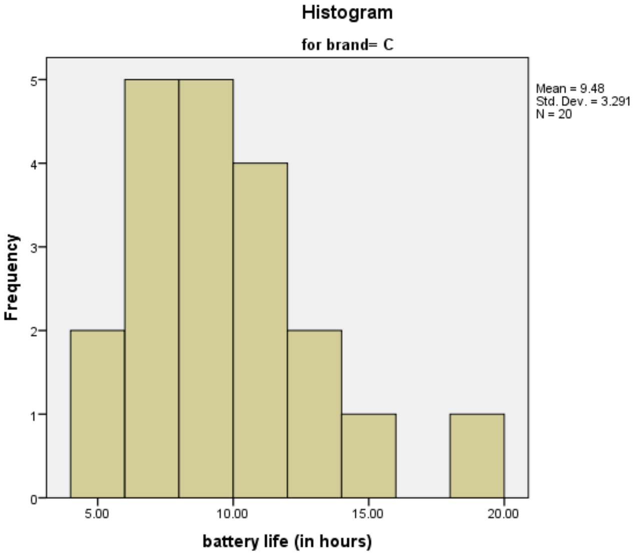
brand

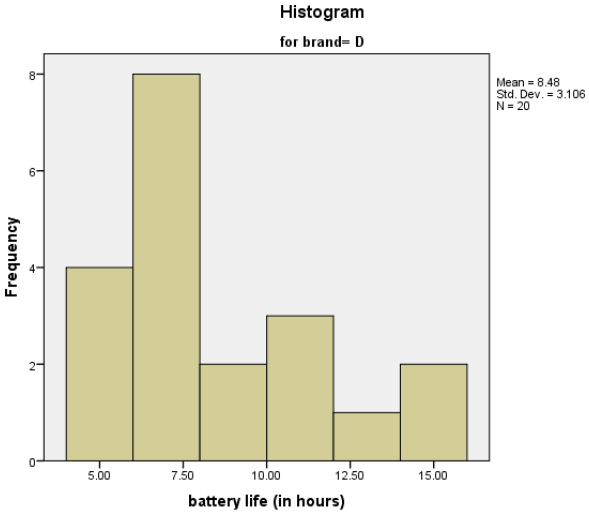
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	20	25.0	25.0	25.0
	B	20	25.0	25.0	50.0
	C	20	25.0	25.0	75.0
	D	20	25.0	25.0	100.0
	Total	80	100.0	100.0	

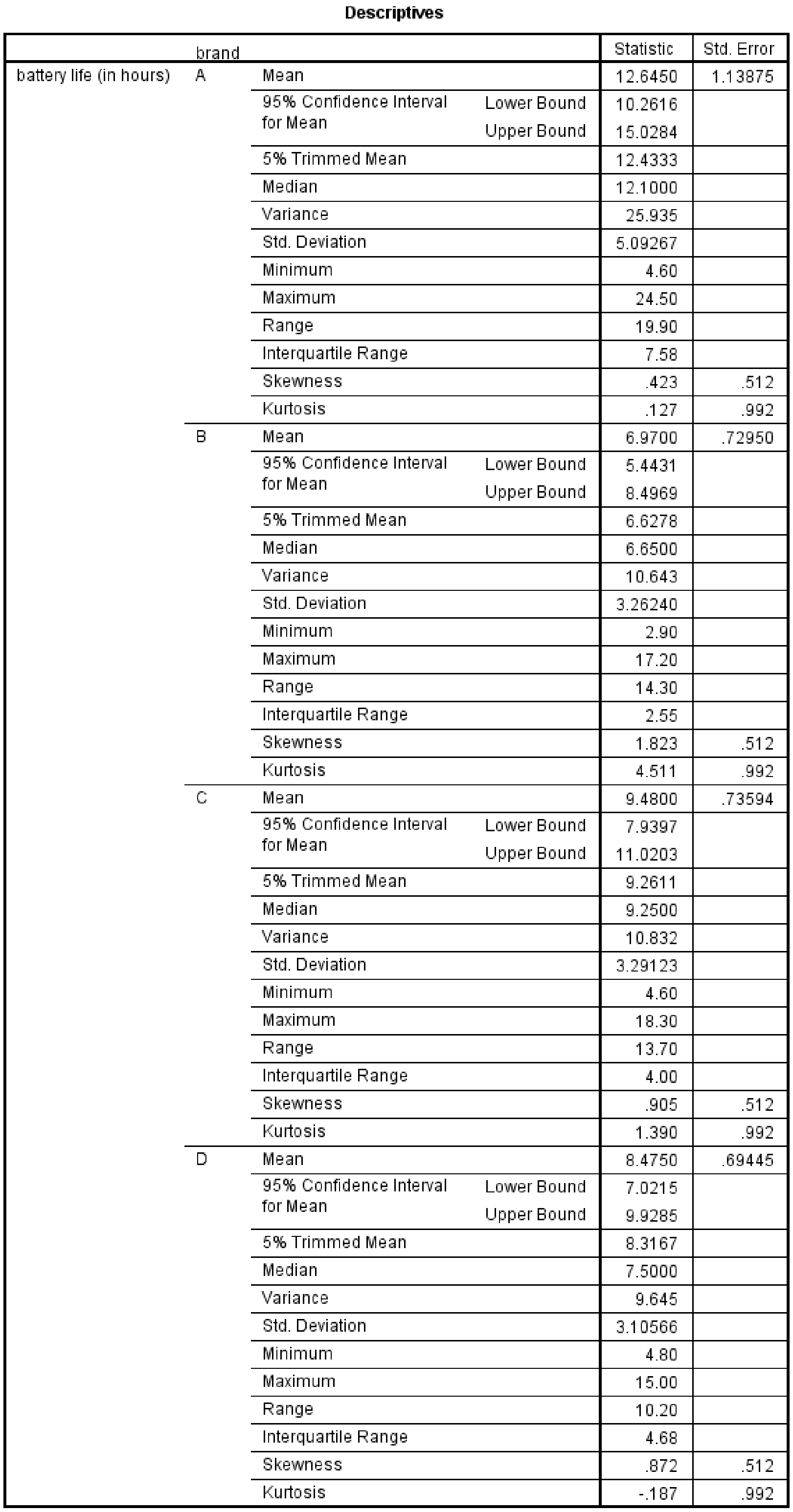



Histogram
for brand= A
Mean = 12.65
Std. Dev. = 5.093
N = 20
Frequency
0
1
2
3
4
5
5.00
10.00
15.00
20.00
25.00
battery life (in hours)


Histogram
for brand= B
Mean = 6.97
Std. Dev. = 3.262
N = 20
Frequency
0
2
4
6
8
10
12
5.00
10.00
15.00
battery life (in hours)


Histogram
for brand= C
Mean = 9.48
Std. Dev. = 3.291
N = 20
Frequency
0
1
2
3
4
5
5.00
10.00
15.00
20.00
battery life (in hours)


Histogram
for brand= D
Mean = 8.48
Std. Dev. = 3.106
N = 20
Frequency
0
2
4
6
8
5.00
7.50
10.00
12.50
15.00
battery life (in hours)



Descriptives

	brand			Statistic	Std. Error
battery life (in hours)	A	Mean		12.6450	1.13875
		95% Confidence Interval for Mean	Lower Bound	10.2616	
			Upper Bound	15.0284	
		5% Trimmed Mean		12.4333	
		Median		12.1000	
		Variance		25.935	
		Std. Deviation		5.09267	
		Minimum		4.60	
		Maximum		24.50	
		Range		19.90	
		Interquartile Range		7.58	
		Skewness		.423	.512
		Kurtosis		.127	.992
	B	Mean		6.9700	.72950
		95% Confidence Interval for Mean	Lower Bound	5.4431	
			Upper Bound	8.4969	
		5% Trimmed Mean		6.6278	
		Median		6.6500	
		Variance		10.643	
		Std. Deviation		3.26240	
		Minimum		2.90	
		Maximum		17.20	
		Range		14.30	
		Interquartile Range		2.55	
		Skewness		1.823	.512
		Kurtosis		4.511	.992
	C	Mean		9.4800	.73594
		95% Confidence Interval for Mean	Lower Bound	7.9397	
			Upper Bound	11.0203	
		5% Trimmed Mean		9.2611	
		Median		9.2500	
		Variance		10.832	
		Std. Deviation		3.29123	
		Minimum		4.60	
		Maximum		18.30	
		Range		13.70	
		Interquartile Range		4.00	
		Skewness		.905	.512
		Kurtosis		1.390	.992
	D	Mean		8.4750	.69445
		95% Confidence Interval for Mean	Lower Bound	7.0215	
			Upper Bound	9.9285	
		5% Trimmed Mean		8.3167	
		Median		7.5000	
		Variance		9.645	
		Std. Deviation		3.10566	
		Minimum		4.80	
		Maximum		15.00	
		Range		10.20	
		Interquartile Range		4.68	
		Skewness		.872	.512
		Kurtosis		-.187	.992

Based on the information provided, work with your assigned group to write an argument that answers the research question.

Expectations:

1. You may discuss this assignment with your assigned group and your instructor only. Any discussion, regardless of how minor, with classmates, tutors, or anyone else will be considered an honor code violation.
2. All group members should be involved in every phase of the work, and all should approve of the final product before it is submitted.
3. Submit your group's typed argument—just one per group with all names on it—to MyCourses in a .doc, .docx, or .pdf format. If you don't trust a group member to submit the assignment on time, you may upload more than one per group, as long as group members turn in the exact same work.
4. Please alert me as soon as possible if any member of your group is not participating in the work.

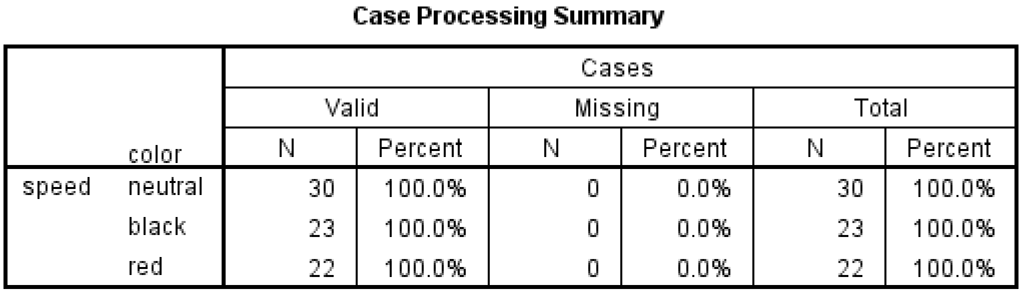
Statistical Reasoning Assignment #2b

Suppose a study is conducted and you are provided with the following information.

Research question: Is there a difference among the average speeds of drivers of neutral-colored cars, red cars, and black cars? (Some people believe that drivers of red and black cars speed more often than other drivers, causing police to watch those drivers more closely.)

Sampling and data collection procedure: A police officer goes to various parts of the same town at various parts of the day and records both the color and speed of different cars as they pass. All the cars were in 60 mph zones when they were clocked.

Results:

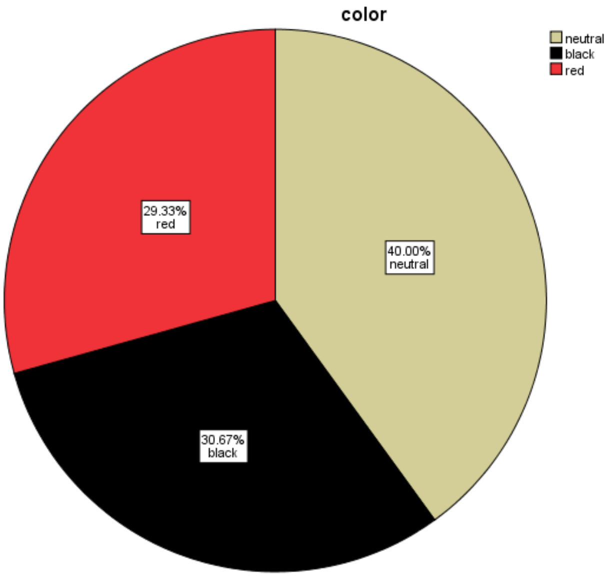


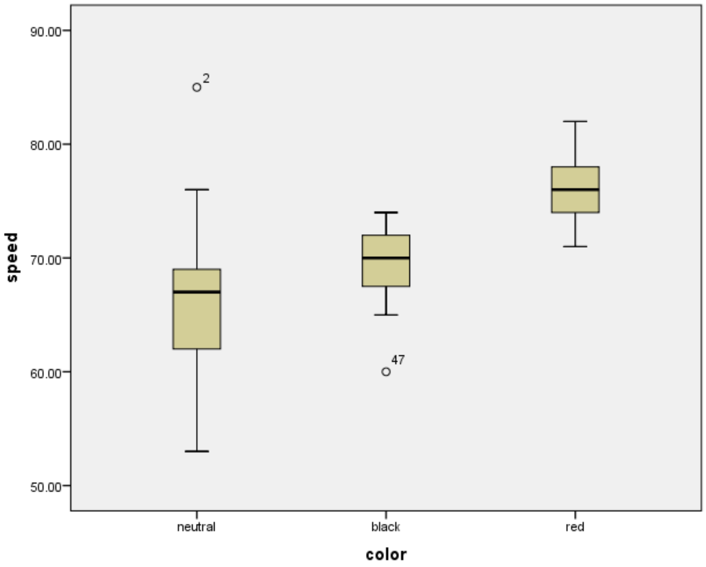
Case Processing Summary

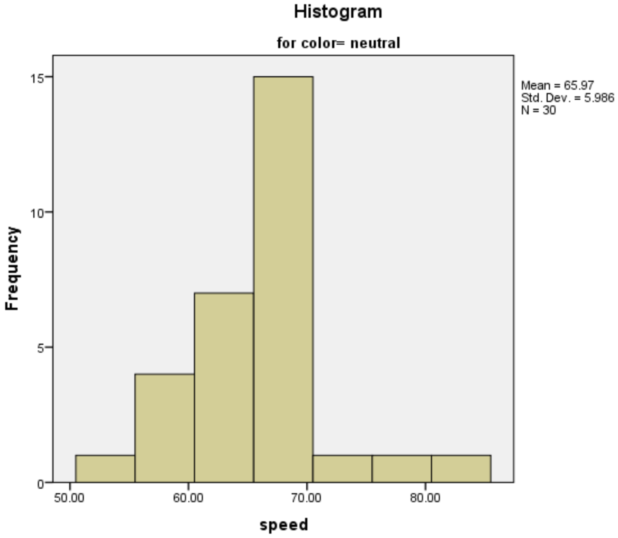
		Cases					
		Valid		Missing		Total	
	color	N	Percent	N	Percent	N	Percent
speed	neutral	30	100.0%	0	0.0%	30	100.0%
	black	23	100.0%	0	0.0%	23	100.0%
	red	22	100.0%	0	0.0%	22	100.0%

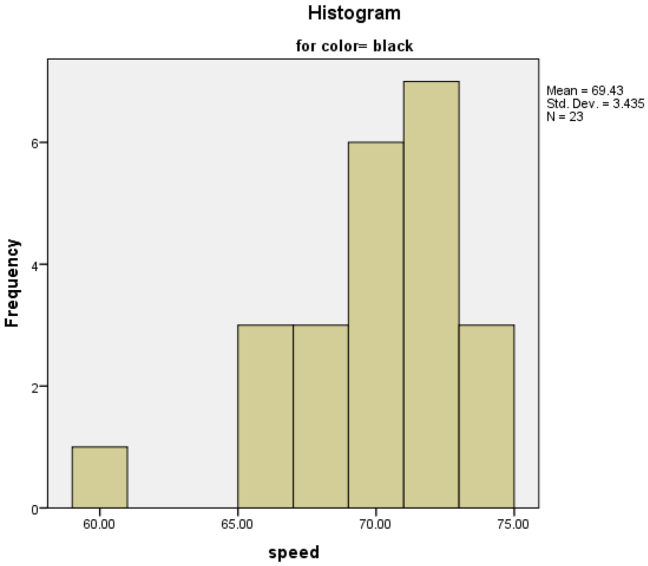
color

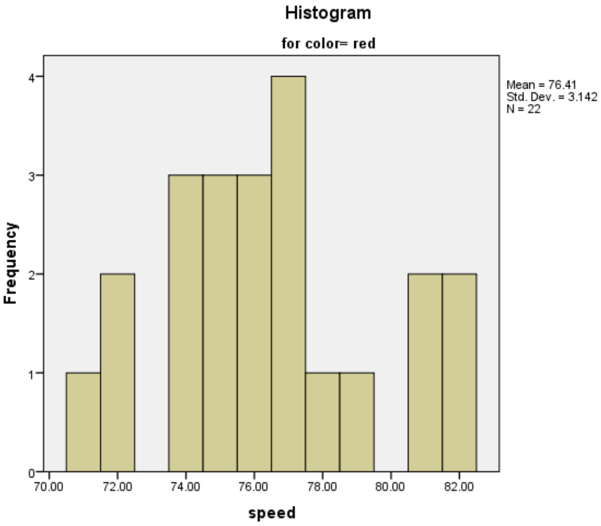
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	neutral	30	40.0	40.0	40.0
	black	23	30.7	30.7	70.7
	red	22	29.3	29.3	100.0
	Total	75	100.0	100.0	

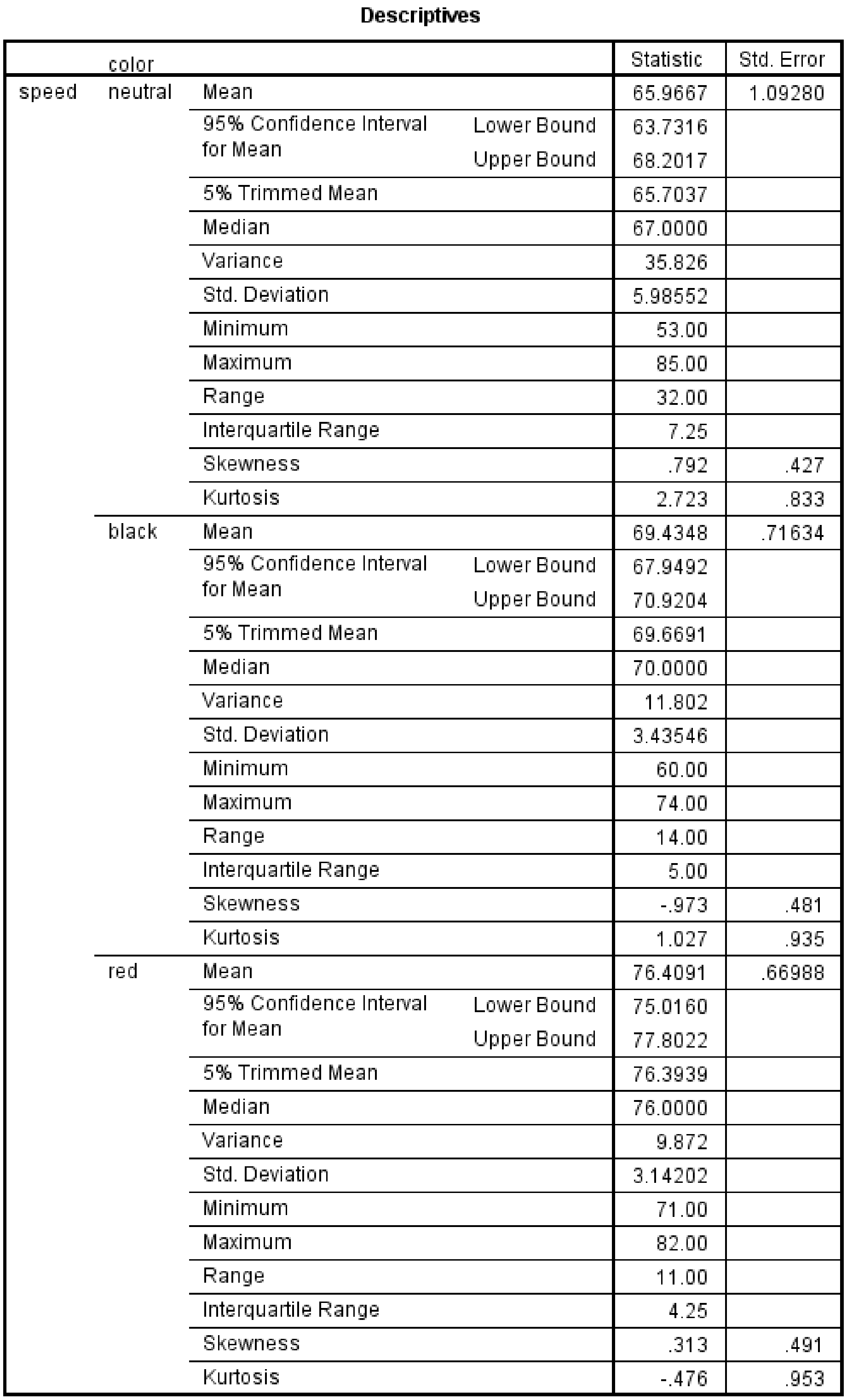



90.00
80.00
70.00
60.00
50.00
speed
2
47
neutral
black
red
color


Histogram
for color= neutral
Mean = 65.97
Std. Dev. = 5.986
N = 30
15
10
5
0
Frequency
50.00
60.00
70.00
80.00
speed


Histogram
for color= black
Mean = 69.43
Std. Dev. = 3.435
N = 23
Frequency
0
2
4
6
60.00
65.00
70.00
75.00
speed


Histogram
for color= red
Mean = 76.41
Std. Dev. = 3.142
N = 22
Frequency
0
1
2
3
4
70.00
72.00
74.00
76.00
78.00
80.00
82.00
speed



Descriptives

	color			Statistic	Std. Error
speed	neutral	Mean		65.9667	1.09280
		95% Confidence Interval for Mean	Lower Bound	63.7316	
			Upper Bound	68.2017	
		5% Trimmed Mean		65.7037	
		Median		67.0000	
		Variance		35.826	
		Std. Deviation		5.98552	
		Minimum		53.00	
		Maximum		85.00	
		Range		32.00	
		Interquartile Range		7.25	
		Skewness		.792	.427
		Kurtosis		2.723	.833
	black	Mean		69.4348	.71634
		95% Confidence Interval for Mean	Lower Bound	67.9492	
			Upper Bound	70.9204	
		5% Trimmed Mean		69.6691	
		Median		70.0000	
		Variance		11.802	
		Std. Deviation		3.43546	
		Minimum		60.00	
		Maximum		74.00	
		Range		14.00	
		Interquartile Range		5.00	
		Skewness		-.973	.481
		Kurtosis		1.027	.935
	red	Mean		76.4091	.66988
		95% Confidence Interval for Mean	Lower Bound	75.0160	
			Upper Bound	77.8022	
		5% Trimmed Mean		76.3939	
		Median		76.0000	
		Variance		9.872	
		Std. Deviation		3.14202	
		Minimum		71.00	
		Maximum		82.00	
		Range		11.00	
		Interquartile Range		4.25	
		Skewness		.313	.491
		Kurtosis		-.476	.953

Based on the information provided, work independently to construct an argument that answers the research question.

Expectations:

1. You may discuss this assignment with your instructor only. Any discussion, regardless of how minor, with classmates, tutors, or anyone else will be considered an honor code violation.
2. Submit your typed argument to MyCourses in a .doc, .docx, or .pdf format.

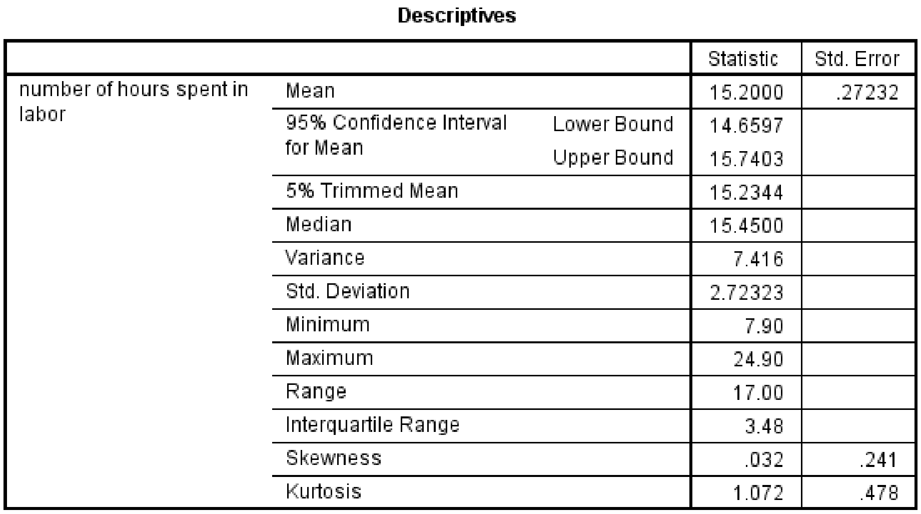
Statistical Argument #3a

Suppose a study is conducted and you are provided with the following information.

Research question: Is a new drug effective in decreasing the amount of time women spend in labor, from the first contraction to delivery? Suppose it is known that the average length of the labor for a first-time mother who has not taken the drug is about 16 hours.

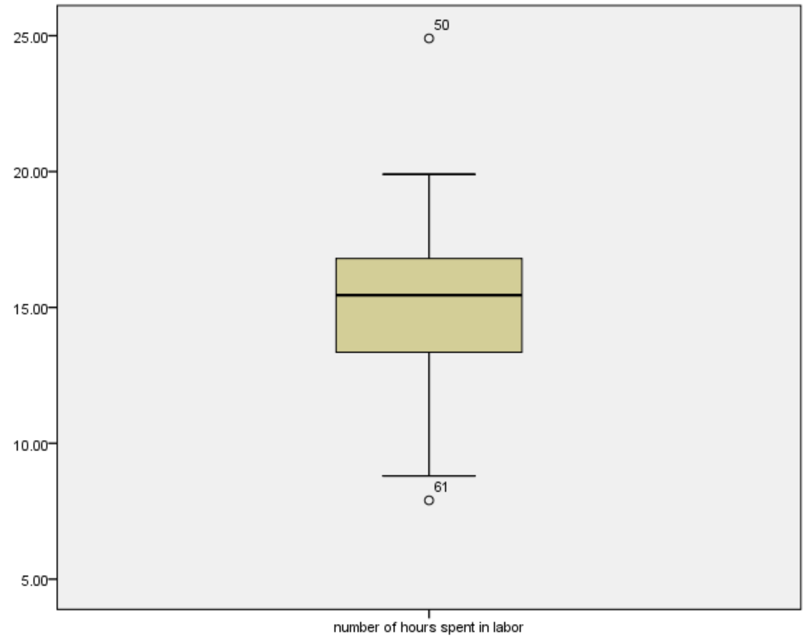
Sampling and data collection procedure: The pharmaceutical company responsible for developing the drug conducted clinical trials using a group of 100 first-time mothers who volunteered for the study. Nurses attending to the patients recorded the time of the first contraction and the time of the birth, then computed the length of the labor.

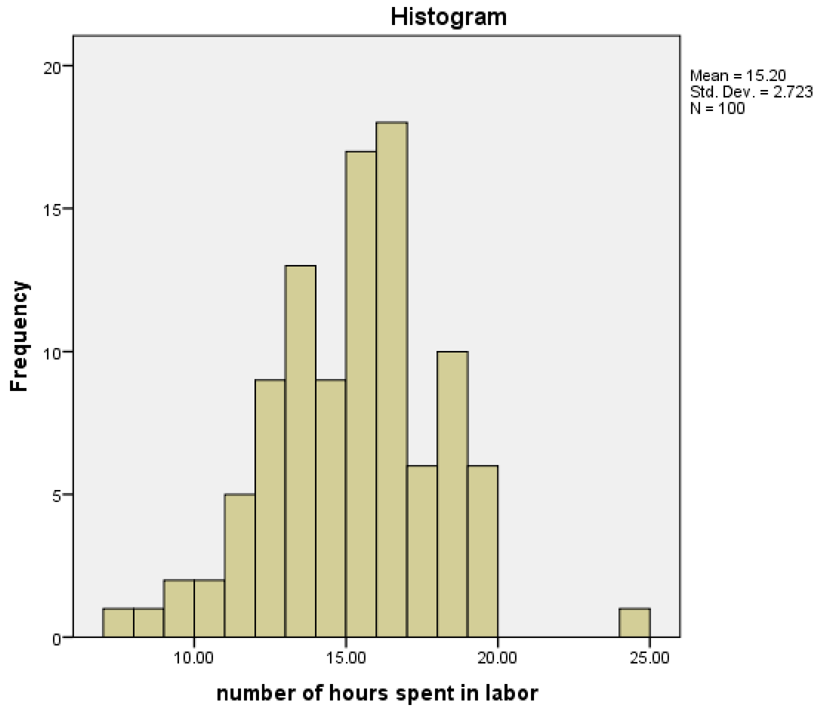
Results:

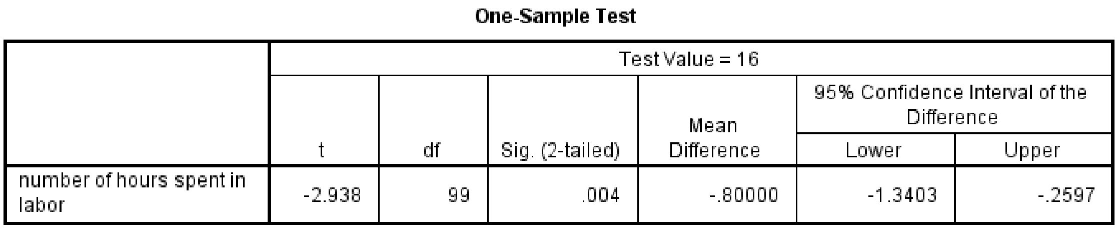


Descriptives

			Statistic	Std. Error
number of hours spent in labor	Mean		15.2000	.27232
	95% Confidence Interval for Mean	Lower Bound	14.6597	
		Upper Bound	15.7403	
	5% Trimmed Mean		15.2344	
	Median		15.4500	
	Variance		7.416	
	Std. Deviation		2.72323	
	Minimum		7.90	
	Maximum		24.90	
	Range		17.00	
	Interquartile Range		3.48	
	Skewness		.032	.241
	Kurtosis		1.072	.478






One-Sample Test

	Test Value = 16					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
number of hours spent in labor	-2.938	99	.004	-.80000	-1.3403	-.2597

Based on the information provided, work with your assigned group to write an argument that answers the research question.

Expectations:

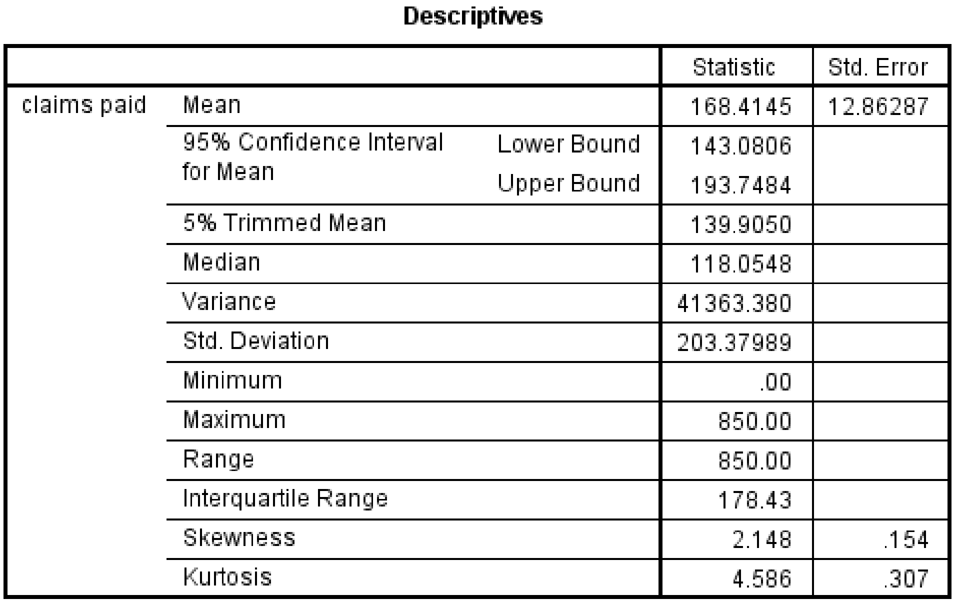
1. You may discuss this assignment with your assigned group and your instructor only. Any discussion, regardless of how minor, with classmates, tutors, or anyone else will be considered an honor code violation.
2. All group members should be involved in every phase of the work, and all should approve of the final product before it is submitted.
3. Submit your group's typed argument—just one per group with all names on it—to MyCourses in a .doc, .docx, or .pdf format. If you don't trust a group member to submit the assignment on time, you may upload more than one per group, as long as group members turn in the exact same work.
4. Please alert me as soon as possible if any member of your group is not participating in the work.

Statistical Reasoning Assignment #3b

Research question: Does an appliance company make a profit when they offer a two-year warranty on a certain brand of refrigerator? Customers pay $175 for the warranty, so if the amount of claims the company must pay to repair or replace the refrigerators is less than $175 on average, they will make a profit.

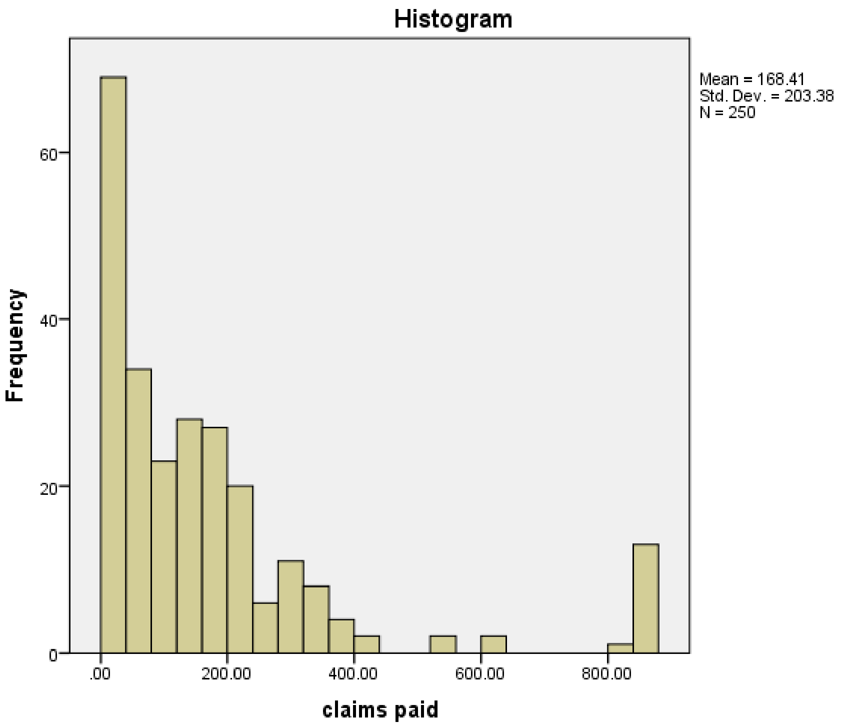
Sampling and data collection procedure: The appliance company uses as their sample the first 250 customers who avail themselves of the offer. They monitor the claims for two years and at the end tally up the total claims.

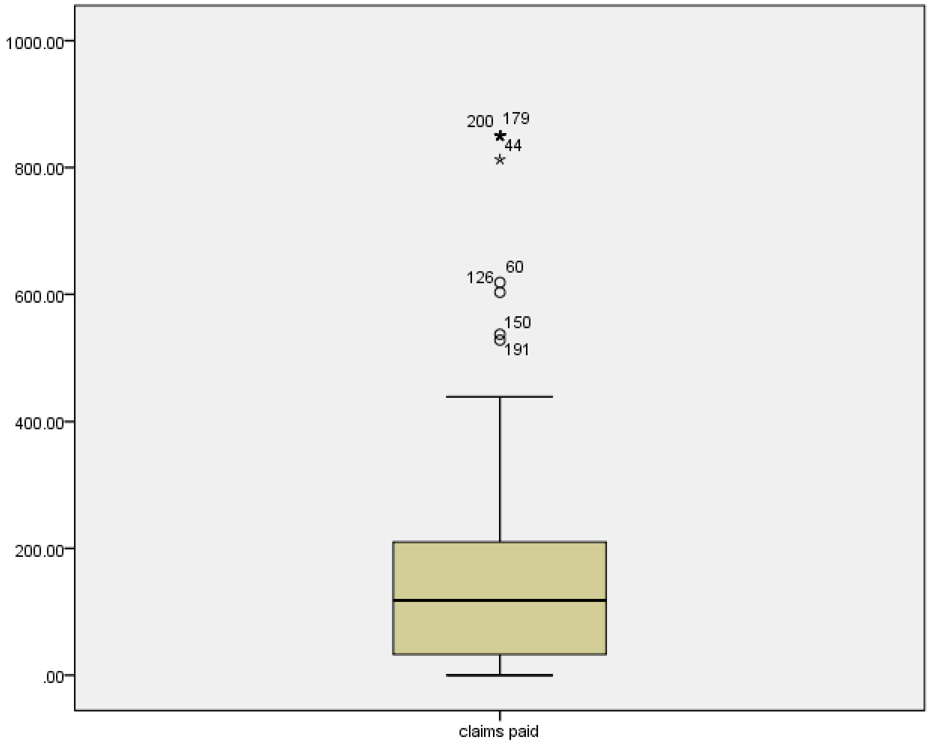
Results:

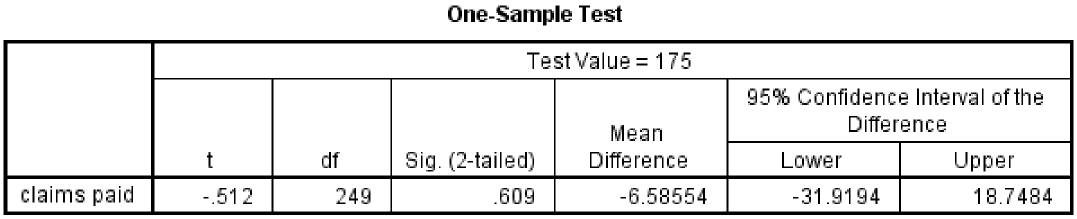


Descriptives

			Statistic	Std. Error
claims paid	Mean		168.4145	12.86287
	95% Confidence Interval for Mean	Lower Bound	143.0806	
		Upper Bound	193.7484	
	5% Trimmed Mean		139.9050	
	Median		118.0548	
	Variance		41363.380	
	Std. Deviation		203.37989	
	Minimum		.00	
	Maximum		850.00	
	Range		850.00	
	Interquartile Range		178.43	
	Skewness		2.148	.154
	Kurtosis		4.586	.307


Histogram
Mean = 168.41
Std. Dev. = 203.38
N = 250
Frequency
0
20
40
60
.00
200.00
400.00
600.00
800.00
claims paid


1000.00
800.00
600.00
400.00
200.00
.00
200
179
44
126
60
150
191
claims paid



One-Sample Test

	Test Value = 175					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
claims paid	-.512	249	.609	-6.58554	-31.9194	18.7484

Based on the information provided, work independently to construct an argument that answers the research question.

Expectations:

1. You may discuss this assignment with your instructor only. Any discussion, regardless of how minor, with classmates, tutors, or anyone else will be considered an honor code violation.
2. Submit your typed argument to MyCourses in a .doc, .docx, or .pdf format.

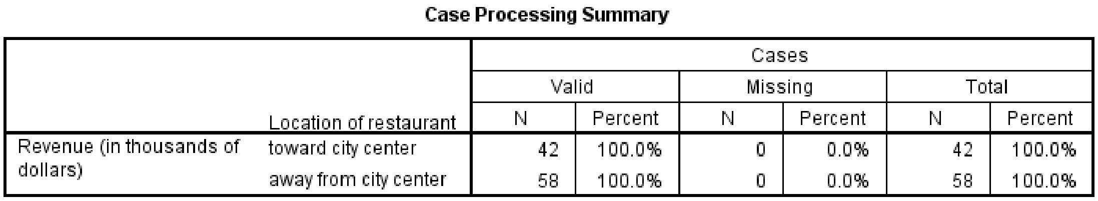
Statistical Reasoning Assignment #4a

Suppose a study is conducted and you are provided with the following information.

Research question: Do fast food restaurants located on the side of the street going toward the city center have higher revenues than fast food restaurants located on the side of the street going away from the city center?

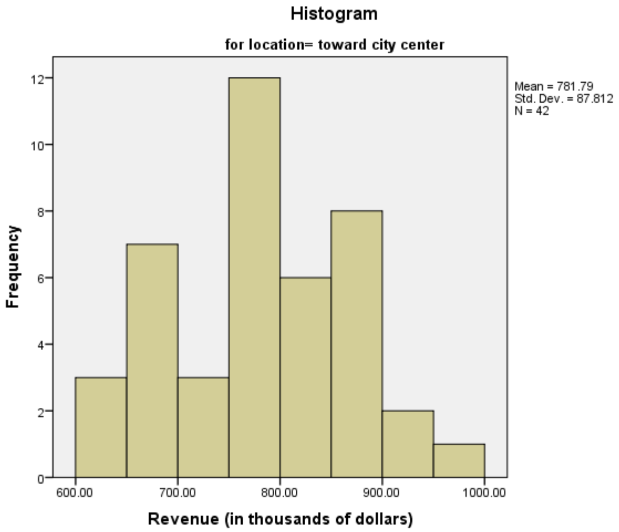
Sampling and data collection procedure: The management of a large fast food chain is trying to determine the optimal location for a future location, so they take a sample of 100 of their top-selling restaurants that are within five miles of city centers. They recorded the location of the restaurants (whether on the side of the street heading toward town or on the side of the street heading away from town) as well as the revenues in the first quarter of this year.

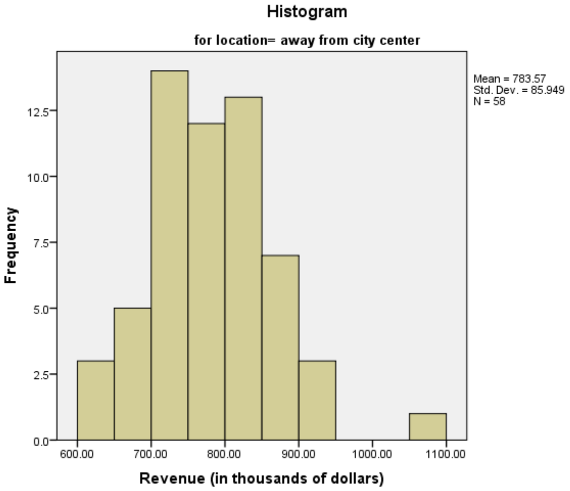
Results:

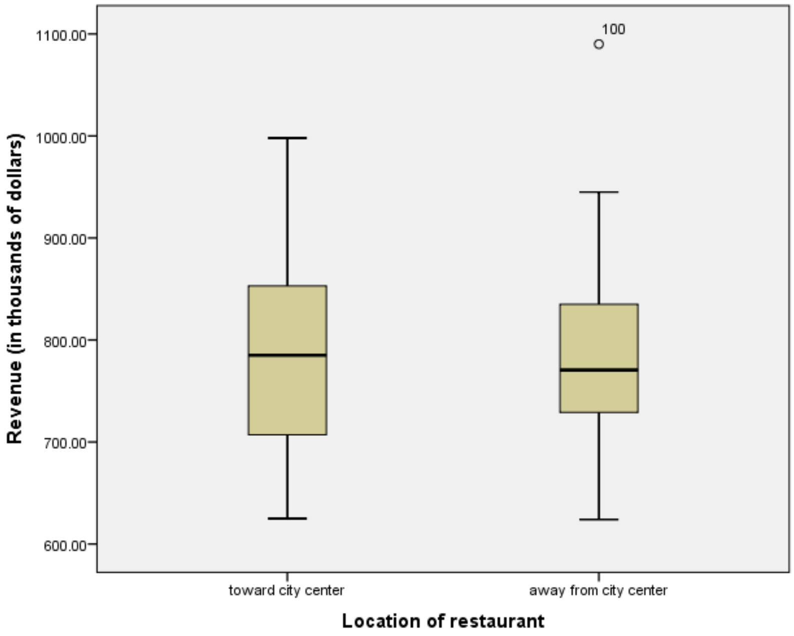


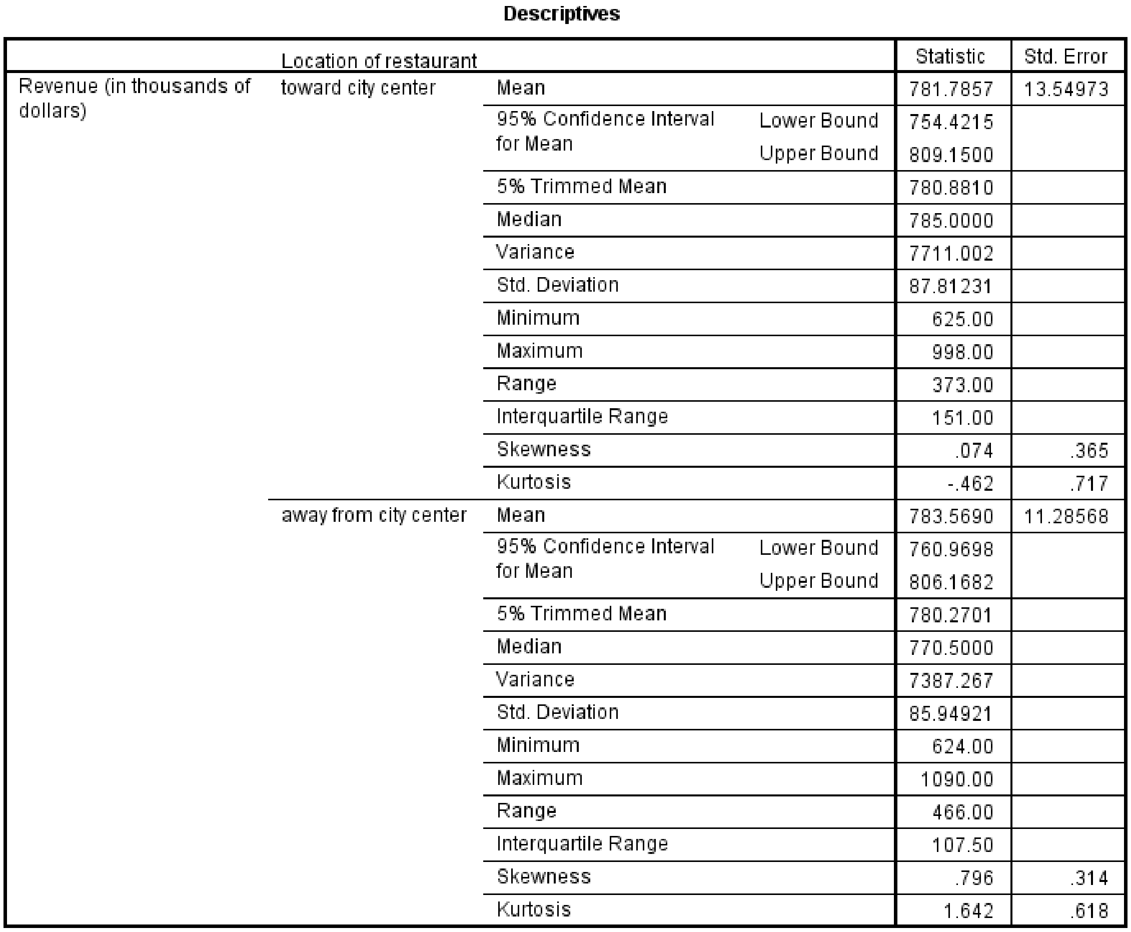
Case Processing Summary

		Cases					
		Valid		Missing		Total	
	Location of restaurant	N	Percent	N	Percent	N	Percent
Revenue (in thousands of dollars)	toward city center	42	100.0%	0	0.0%	42	100.0%
	away from city center	58	100.0%	0	0.0%	58	100.0%

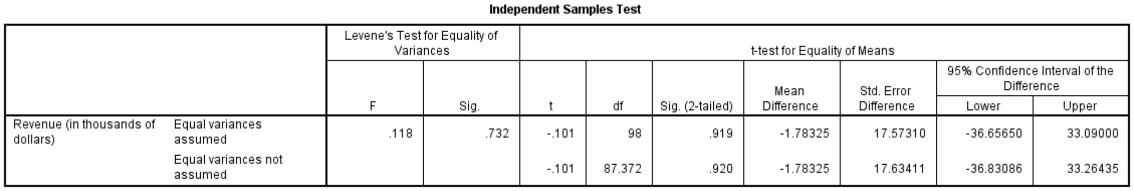

Histogram
for location= toward city center
Mean = 781.79
Std. Dev. = 87.812
N = 42
Frequency
0
2
4
6
8
10
12
600.00
700.00
800.00
900.00
1000.00
Revenue (in thousands of dollars)


Histogram
for location= away from city center
Mean = 783.57
Std. Dev. = 85.949
N = 58
Frequency
0.0
2.5
5.0
7.5
10.0
12.5
600.00
700.00
800.00
900.00
1000.00
1100.00
Revenue (in thousands of dollars)




Descriptives

	Location of restaurant			Statistic	Std. Error
Revenue (in thousands of dollars)	toward city center	Mean		781.7857	13.54973
		95% Confidence Interval for Mean	Lower Bound	754.4215	
			Upper Bound	809.1500	
		5% Trimmed Mean		780.8810	
		Median		785.0000	
		Variance		7711.002	
		Std. Deviation		87.81231	
		Minimum		625.00	
		Maximum		998.00	
		Range		373.00	
		Interquartile Range		151.00	
		Skewness		.074	.365
		Kurtosis		-.462	.717
	away from city center	Mean		783.5690	11.28568
		95% Confidence Interval for Mean	Lower Bound	760.9698	
			Upper Bound	806.1682	
		5% Trimmed Mean		780.2701	
		Median		770.5000	
		Variance		7387.267	
		Std. Deviation		85.94921	
		Minimum		624.00	
		Maximum		1090.00	
		Range		466.00	
		Interquartile Range		107.50	
		Skewness		.796	.314
		Kurtosis		1.642	.618



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Revenue (in thousands of dollars)	Equal variances assumed	.118	.732	-.101	98	.919	-1.78325	17.57310	-36.65650	33.09000
	Equal variances not assumed			-.101	87.372	.920	-1.78325	17.63411	-36.83086	33.26435

Based on the information provided, work with your assigned group to write an argument that answers the research question.

Expectations:

1. You may discuss this assignment with your assigned group and your instructor only. Any discussion, regardless of how minor, with classmates, tutors, or anyone else will be considered an honor code violation.
2. All group members should be involved in every phase of the work, and all should approve of the final product before it is submitted.
3. Submit your group's typed argument—just one per group with all names on it—to MyCourses in a .doc, .docx, or .pdf format. If you don't trust a group member to submit the assignment on time, you may upload more than one per group, as long as group members turn in the exact same work.
4. Please alert me as soon as possible if any member of your group is not participating in the work.

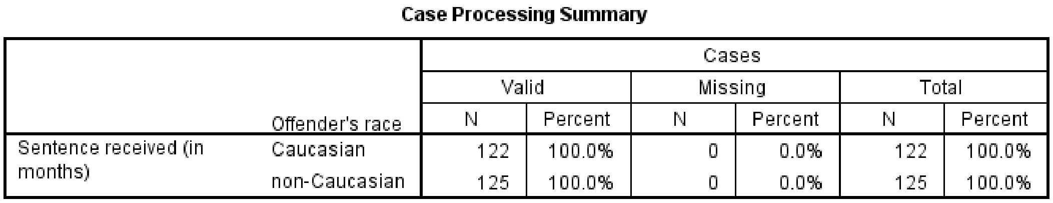
Statistical Reasoning Assignment #4b

Suppose a study is conducted and you are provided with the following information.

Research question: Is there racial bias in sentencing for felonies?

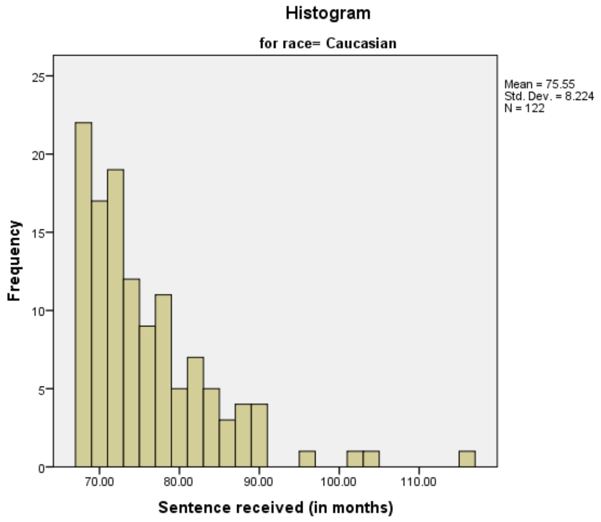
Sampling and data collection procedure: A criminologist searches a database of criminal offenses in a given state, and downloads records for all offenders convicted of burglary in the previous year. The sample consists of all the burglaries for that year. The criminologist notes the length of the sentence (in number of months) and the offender's race. Based on a prior study, he believes Caucasian offenders will get more lenient sentences than any other group, so he categorizes the offenders as either Caucasian or not Caucasian.

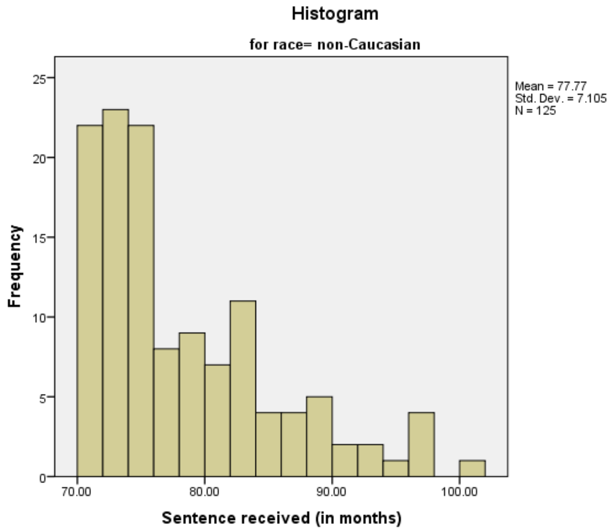
Results:

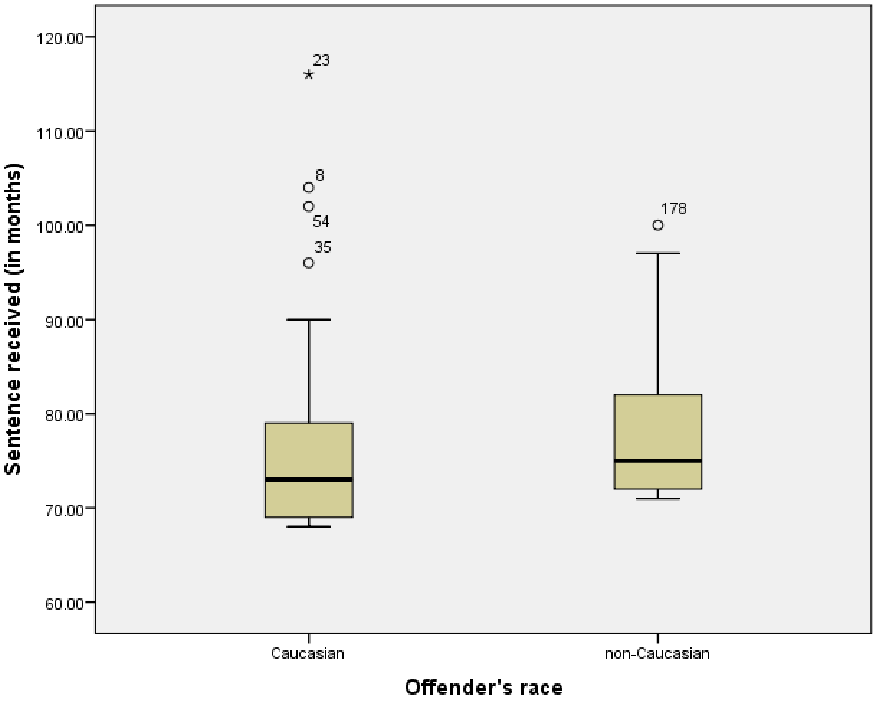


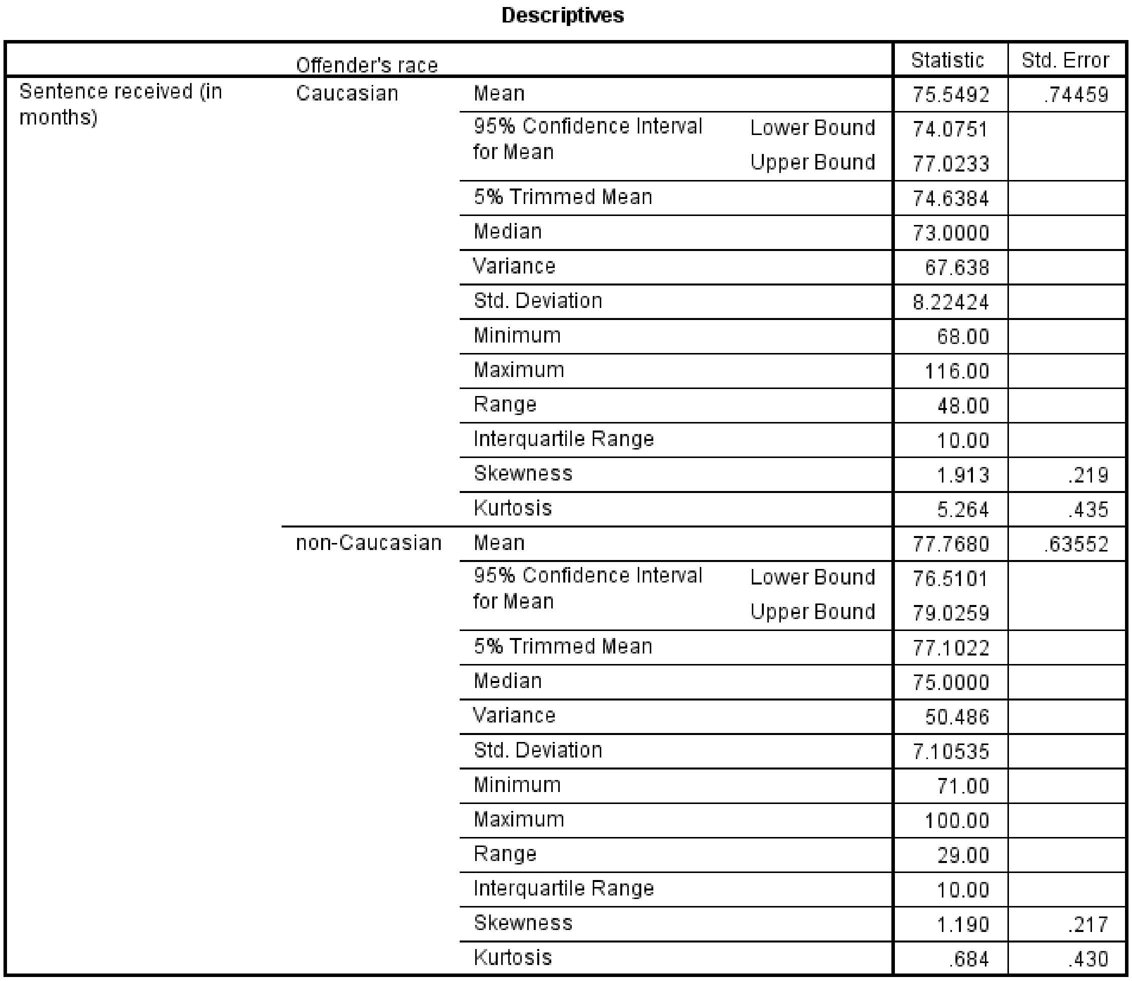
Case Processing Summary

		Cases					
		Valid		Missing		Total	
	Offender's race	N	Percent	N	Percent	N	Percent
Sentence received (in months)	Caucasian	122	100.0%	0	0.0%	122	100.0%
	non-Caucasian	125	100.0%	0	0.0%	125	100.0%


Histogram
for race= Caucasian
Mean = 75.55
Std. Dev. = 8.224
N = 122
Frequency
0
5
10
15
20
25
70.00
80.00
90.00
100.00
110.00
Sentence received (in months)

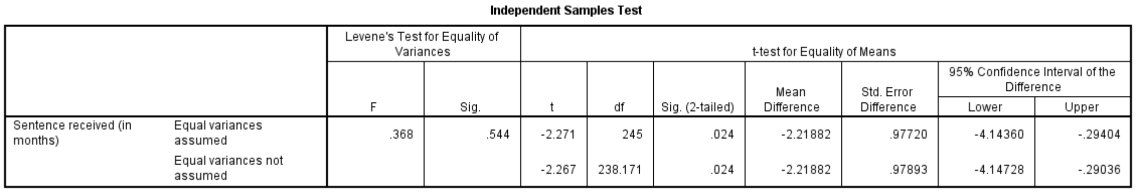

Histogram
for race= non-Caucasian
Mean = 77.77
Std. Dev. = 7.105
N = 125
Frequency
0
5
10
15
20
25
70.00
80.00
90.00
100.00
Sentence received (in months)


120.00
110.00
100.00
90.00
80.00
70.00
60.00
Sentence received (in months)
23
8
54
35
178
Caucasian
non-Caucasian
Offender's race



Descriptives

	Offender's race			Statistic	Std. Error
Sentence received (in months)	Caucasian	Mean		75.5492	.74459
		95% Confidence Interval for Mean	Lower Bound	74.0751	
			Upper Bound	77.0233	
		5% Trimmed Mean		74.6384	
		Median		73.0000	
		Variance		67.638	
		Std. Deviation		8.22424	
		Minimum		68.00	
		Maximum		116.00	
		Range		48.00	
		Interquartile Range		10.00	
		Skewness		1.913	.219
		Kurtosis		5.264	.435
	non-Caucasian	Mean		77.7680	.63552
		95% Confidence Interval for Mean	Lower Bound	76.5101	
			Upper Bound	79.0259	
		5% Trimmed Mean		77.1022	
		Median		75.0000	
		Variance		50.486	
		Std. Deviation		7.10535	
		Minimum		71.00	
		Maximum		100.00	
		Range		29.00	
		Interquartile Range		10.00	
		Skewness		1.190	.217
		Kurtosis		.684	.430



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Sentence received (in months)	Equal variances assumed	.368	.544	-2.271	245	.024	-2.21882	.97720	-4.14360	-.29404
	Equal variances not assumed			-2.267	238.171	.024	-2.21882	.97893	-4.14728	-.29036

Based on the information provided, work independently to construct an argument that answers the research question.

Expectations:

1. You may discuss this assignment with your instructor only. Any discussion, regardless of how minor, with classmates, tutors, or anyone else will be considered an honor code violation.
2. Submit your typed argument—just one per group with all names on it—to MyCourses in a .doc, .docx, or .pdf format.

APPENDIX C: FULL TEXT OF AMBER’S STATISTICAL ARGUMENTS

TASK 1

After analyzing the Statistical Reasoning Assignment #1B, I found numerous errors throughout the study. The research question the survey provided wonders if, “Teens pay attention to local and national politics”. Instead of the survey asking if teens pay attention to local and national politics, they should revise the question and state, “Do teenagers age (13-19) watch the local news, and stay aware of the current events?” This way the survey would pertain to teenagers from ages 13 to 19. Also, the question would ask if they are staying up to date with current news.

The study sampled 50 teens that were recruited from a popular arcade at a popular mall. The study only sampled 50 teenagers in one specific area which is defective to the study in numerous ways. First, 50 teenagers does not represent teenagers across the nation. Second, the sample only consisted of teenagers that were located at a popular arcade at the most popular area in the city. Sampling only teenagers at an arcade in the mall does not make the study valid because the teens are only in one location. The sample should have asked teenagers in different locations such as the library, school, park, movie theatre, or food court. Also, the sample should have been at least 50 teenagers from different states or different cities to keep a random sample of teens across the world.

The sample question the survey asked the teens should have been easier for a middle school or high school student to answer. Asking the question “Would you support suffrage for 16-and-17 year olds?” would be confusing for a teenager to answer and would lead to an answer such as No. The sample should have ask a question that they feel would interest a teenagers mind, such as “Would you support cyberbullying?” A question like this would interest a teenager more because they possibly would have heard about it through friends, family, and most likely seen certain events on the news when teenagers have killed themselves because of the effect of bullying.

The survey the organizers developed was done in an insupportable way. I believe an accurate poll would be more effective if they asked teenagers in different areas around the world and numerous locations. Also, if they changed the question they are asking the teenagers to a question that would interest a teenager from the ages 13-19. This would enable the teenagers to know the question they are responding to and the poll would be represented more accurately then only being in one location.

TASK 2

After analyzing the Statistical Reasoning Assignment 2B, I found a few errors in the research question and sampling data collection procedure for the research. The scenario of the research I reviewed made sense. However, I feel as though the research question is difficult to interpret, instead the research question should be revised and state “Do drivers of certain colored vehicles such as neutral-colored, black, or red cars speed more often”. This way the study is based on the vehicle color and it is easier to interpret. I also stated that for the sampling data procedure, police officers should have recorded vehicles speeding and the specific color of their vehicle in multiple towns instead of the same town. This way the study would be more beneficial because the study would be various between each town and possibly include more vehicles.

The speed of neutral-colored vehicles has a mean of 65.9667 units. The speed of black-colored vehicles has a mean of 69.4348 units. The speed of red-colored vehicles has a mean of 76.4091 units. Throughout the research the neutral-colored vehicles had a higher percentage of speeding shown in the results on the pie-chart. The neutral-colored vehicles had a 40% chance of speeding, while the black-colored vehicles had a 30.67% and red-colored vehicles 29.33%. The study also included more neutral-colored vehicles in the research than the red and black one. In the relative frequency distribution chart there were 30 neutral-colored vehicles, 23 black-colored vehicles and 22 red-colored vehicles. The standard deviation for the vehicles also varied between each one with the neutral vehicles having the highest standard deviation of 5.98552 units since the sample size was larger, the black vehicles had a standard deviation of 3.43546 units and red vehicles had a standard deviation of 3.14202 units. By looking at the neutral-colored vehicle boxplot, I can see that it has an outlier of 2 units and the black-colored vehicle boxplot has an outlier of 47 units. The min and max for each vehicle are relatively different units for each one, neutral-colored vehicles has a min of 53.00 units and max of 85.00 units, black vehicles have a min of 60.00 units and max of 74.00, and red vehicles have a min of 71.00 units and 82.00 units. The histogram for neutral-colored vehicles represents a bell-shape curve, the black-colored vehicles represent a left skewed shaped and the red-colored vehicles represent a bimodal shape. The interquartile range on the boxplot for the neutral-colored vehicles spreads out the most to 7.25 units, the black vehicles spread out 5.00 units, and red vehicles 4.25 units. Although the research had a few errors with the research question and sampling data collection procedure, the study still included reasonable data to research and provided accurate charts to review for the study.

TASK 3

After analyzing the Statistical Reasoning Assignment 3B, I found a few errors in the research question and data collection series for the research. The research question is plausible, and I feel as though it provides reasonable data to receive accurate answers for the study. However, I feel as though the research should name the specific type of brand of refrigerators they are going to receive the information on, for if the company makes a profit when they offer a two-year warranty. To improve the research question, it should be rephrased and state, “Does an appliance company make a profit when they offer a two-year warranty on a specific brand of refrigerator such as Frigidaire”. This way the study will be based on one specific brand of refrigerator and the sampling and data collection procedure will change to only sample the amount of customers that purchased a Frigidaire refrigerator. For the measure of center for the data I used the mean, which was $168.4145. The spread of the data is $203.3798, which is the standard deviation. The distribution is right-skewed, and there are multiple observations numbers. There is an observation of 191, 150, 126, and 60 at 500.00 to 600.00 claims paid and 44,179,200 at 800.00 to 900.00 claims paid. For the t-test, I used a sample of 250 customers who brought a warranty for their refrigerator. The possible hypotheses are Ho: u=175, H1: u<175. If the sample mean is less than the hypothesized value, the alternative hypothesis will be u<175. The p-value for the test is .3045. The test will reject Ho if p-value is less than 5% level of significance. The study Fails to reject Ho. There is not sufficient evidence to conclude that the amount of claims the company must pay to repair or replace the refrigerators is less than $175 on average, and they will make a profit. It is a small effect of .0324, and the guidelines are .2, .5 and .8.

Although the case study included a few errors such as the research question and how the study was sampled, it still provided reasonable data. As I stated before, if the study rephrases the research question to be able to include only the specific brand of refrigerators, the study would be more accurate and provide better data for the research. The assumption to run the t-test has been met. However the study did not provide sufficient amount of evidence to conclude that the company will make a profit if the amount of claims the company must pay to repair or replace the refrigerators is less than $175.

TASK 4

After analyzing the Statistical Reasoning Assignment 4B, I found a few errors with the sampling and data collection procedure. The research question is plausible, and makes logical sense. The research question is asking if “there is a racial bias in sentencing for felons?” However, the sampling and data collection procedure does not yield a representative sample because the sample only consisted of people that convicted a felony of burglary. The sampling and data collection procedure should consist of all convicts that committed a felony, and then sample that data by race. With a sample of all convicts that committed a felony there would be a larger sample to research.

Group 1 for the research is Caucasian offenders convicted for burglary, and Group 2 is Non-Caucasian offenders convicted for burglary. The total sample for the groups of people that were convicted of burglary in the past year is 247; Group 1 has a sample size of 122 Caucasians and Group 2 has a sample size of 125 Non-Caucasians.

Since both groups have a large sample size the assumptions for us to conduct a hypothesis test have been met. For the measure of the center of data I used the mean which is 75.5492 months for Group 1, and 77.7680 months for Group 2. The spread of the data is 8.22424 months for Group 1, and 7.10535 months for Group 2, which is the standard deviation. The histograms for both groups represent a right-skewed distribution. The boxplot for Group 1 has four outliers, 34 and 54 are between 95-100 months, 8 are between 100-110 months, and 23 are between 110-120 months. The boxplot for Group 2 has one outlier of 178 between 100-110 months. The possible hypotheses are Ho: $\mu_1 = \mu_2$, H_1: $\mu_1 > \mu_2$. If the mean for Group One is greater than the mean for Group Two, the alternative hypothesis will be $\mu 1 > \mu 2$. The p-value for the test is .012. The test will reject Ho if the p-value is less than 5% level of significance. The study will reject Ho. There is statistically significant enough evidence to conclude that Caucasian offenders will get more lenient sentences than any other group. The data is a small effect of .2698; the guidelines are .2, .5, and .8. A small effect size means that there is hardly any difference between the data we have been given for this research.

Even though I feel like the research sample and data collection procedure could change, the research still included enough evidence to conclude that Caucasian offenders will get more lenient sentencing than any other group.

APPENDIX D: FULL TEXT OF LEAH'S STATISTICAL ARGUMENTS

TASK 1

The study in the statistical reasoning assignment 1b shows several fundamental flaws in the structure and in its procedural gathering of the data.

The research question, “Do teens pay attention to local and national politics?” is an interesting inquisition, but for statistical purposes, the organizer of the study needs to be more specific in regards to the population. The way the question is worded could mean the population, is all the teens in the world, which is way too broad of a study. This should be narrowed down to teens in the United States. This way simple random sampling can be constructed using a sample of teens here in the U.S. This enhances the accuracy of the data and there is no question where the data was retrieved.

The teenagers that were selected for the sample were 50 recruited students from a popular arcade at a local mall. Again, we have no idea where the study is taking place. Locations of surveys are extremely important to ensure that random simple sampling is being used and that the sample is representing the population of the study. The amount of students in this survey is not large enough to represent the population. The organization conducting the study should have pick one large public high school from each division of all the regions of the U.S. A poll from each high school should consist of at least 250 randomly selected 16 to 17 year old students. (For example, Survey from Northeast Region; Division 1- New England -one large public high school of 250 students, between ages 16 and 17). This procedure would strengthen the representation of the population of all teens in the U.S. Also, another flaw with the sample is the actual recruitment of the respondents. The recruitments could have been selected by the organizer and could lead to biases in the study. Participants of surveys should be randomly selected.

The survey question: “Would you support suffrage for 16 and 17 year olds?” encompasses several errors. For one, I don’t think many 16 or 17 year olds know what the term suffrage means. The organizer assumes the teenager is familiar with political terms. This would definitely skew the participant’s response to not being accurate or not responding at all, because they don’t understand the question that is being asked. The data that was collected makes me believe that this was true. Seventy six percent of the participants responded no to their right to vote. This is not reasonable to me. The teenagers I know want every opportunity to be treated as an adult. I would think more participants would have voted yes to supporting suffrage. A survey question should be asked in a way that the majority of the participants would be able to understand, so an accurate data collection can be obtained. Secondly, the survey question itself is not a good representation of the research question being asked. The data collected from this question is only letting us know what teens think of voting not necessarily that they pay attention to politics. A better question could have been formulated to correlate with the research question, such as: “Do you read or watch local and national news weekly?” The data that would be collected from this question would let the public know that 16 and 17 years olds are either paying attention or not paying attention to their local or national news, which consistently includes political issues.

The final problem with this study is that the public has no idea who is conducting the study or what organization is actually giving the survey. This creates a problem with the validity of the study. The public is less likely to acknowledge a study that is done by an anonymous organizer and that is not supported by a reputable polling organization. This leads the public to infer that biases could be involved, which devalues the validity of the entire study.

As we can see this study has numerous flaws, but with proper sampling and data collection procedures, the study could be sound and useful to the public.

TASK 2

The Research Question “Is there a difference among the average speeds of drivers of neutral-colored cars, red cars, and black cars?” and the data collected do correspond, but unfortunately the data collection procedure is flawed in several different ways, which makes the data results unreliable and flawed.

This observational study was performed by a police officer that clocks people every day for speeding. Yes, he is an expert in the field, but he could be a biased observer for the study. It even states in the Research Question, “some people believe that drivers of red and black cars speed more often than other drivers, causing **Police Officers** to watch those drivers more closely”. This study should be conducted by a reputable trained professional that has no correlation to the law or the people conducting the study. Also, the study needs to be conducted in several different cities to ensure a more random sampling of the car speeds. Cities vary in their tolerances of car speeds and populations vary, which could affect speeds because of traffic. A random sampling of cities all over the U.S. would depict an accurate study. A sampling of neutral, red, and black cars and their speeds from each region of the U.S. would be an example of a more reliable study. The study could have also used a larger sample size for each car. The sample size of the neutral color car was 30, the other samples should have had at least 30 cars in their sample size. The lower number of cars in the sample size leads me to believe that the officer got the numbers he wanted, so he didn’t wait to get anymore red colored vehicles or black, he just stopped at random numbers. The data would have been more reliable if the sample sizes were equal in number to compare.

When viewing the Histograms, I noticed there were two skewed and one normal distribution, so I decided to use the median to find the center. The median for the neutral colored cars is 67.00 mph. The median for black cars is 70.00 mph and the median for red cars is 76.00 mph. Neutral colored cars have the lowest median mph with 67.00 and red colored cars have the highest median mph with 76.00. Each colored car category of the standard deviations were described and based on the statistics the neutral colored cars had the highest variability with their speed at 5.9855 mph and red colored cars had lowest variability at 3.1420 mph. The black colored cars standard deviation was 3.4355 mph. This is unusual to me because normally the bigger the sample size the less variability. In this case, the neutral cars are the largest sample and have the highest variability. I feel that the samples should have been larger and all the same size to have less variability. Also, depending where these cars were clocked, could have had a dramatic effect on how fast a car travels. This could explain some variability. For example, the officer decided to clock red cars in a specific area where people feel more comfortable speeding or more likely to speed (example: going down a hill in a less populated area where police don’t patrol and the afternoon getting off work).

The distributions of these three categories are varied with the black cars having a normal distribution and the red and neutral cars being skewed. The neutral colored cars have a high outlier of 85 mph, the black cars having a low outlier of 60 mph and the red cars showing no outliers. According to the knowledge we have of this study, the police officer could have clocked a neutral colored car in one area and then clocked 20 neutral cars in another area. This could explain why we have one outlier of 85 mph in the neutral colored car category.

As we can see, the data that was collected relates to the research question but the data collection procedure has many problems resulting in unreliable and flawed data. A reputable non-biased trained professional should have done this observational study. This would have admonished any doubts in how the data results were collected. By using the officer, brings up questions of whether he purposely clocked red cars in known unpatrolled areas where people might be more likely to speed to skew the data. Also, this study needs random sampling in different regions all over the U.S. to accurately study the population. The sample sizes should have been larger and all the same, which would make for less variation in the samples to give us more accurate data. Lastly, I feel more information should have been given to us referring to locations clocking the speed of the vehicles. This would have left less room for questionable data.

TASK 3

The research question, “Does an appliance company make a profit when they offer a two – year warranty on a certain brand of refrigerator?” and the data that was collected, corresponds with the study being performed. Unfortunately, the data collection procedures are flawed which makes the study’s results questionable.

This study uses a sample of the first 250 customers, a convenience sample. This is not particularly bad because the company is wanting to use the data for their own benefit (we're assuming), so trying to get data from another location or some other way would not be beneficial. The appliance company conducted their own study, we're assuming that they were skilled in this type of data collection and if so each customer that bought that brand of refrigerator was offered the warranty and was properly informed of the details. But, if the appliance company really wanted to ensure the most accurate results, and ensure that the study had no biases, using an outside reputable professional survey company would have provided that security. Incidents do happen, for example: when a company uses an internal source to do accounting work, that employee could become mad at the company and decide to mess with the numbers to create a loss for the company. I know this sounds far-fetched, but these situations do occur. In this study, the data's accuracy is extremely important for the company's ability to ensure profitability. Which brings me to the data collection process of using one brand of refrigerator but not categorizing the models. I feel this may have flawed the data results with too much variation between the models. Some models are low end, which would more than likely have more problems and there would be price differences in the parts used which would affect the claims paid. Higher end models would probably have fewer problems but parts might be more expensive. This situation might explain the number of outliers. It would be helpful to the company to see which models were the ones giving customers the most problems and which ones didn't. This could help the company to determine which models they were making a profit on and which ones they weren't. In the end, they could decide not to carry the warranty on certain models or just not sell those models in their store.

This study does use appropriate units to describe the data, using claims paid (dollars). The T- test can be used because the sample size is larger than 30, even though the histogram shows a right skewed distribution not a normal distribution. The study shows a mean of 168.4145 dollars in claims paid and a standard deviation of 203.37989 dollars in claims paid. According to the box plot, the maximum data value is 450.00 dollars in claims paid and the minimum data value is 00 dollars in claims paid. There are 18 outliers shown on the Histogram and box plot. There are 15 outliers at 850.00 dollars in claims paid and 2 outliers at 600.00 dollars claims paid and 2 outliers at 550.00 dollars claims paid. This large number of outliers tells us that this is just more than an error. It would be wise for the company to include these outliers in the sample data and then again with the outlier excluded, so that they can tell what effect the outliers have on the results.

After completing a hypotheses test, my results show Ho: u= 175 dollars in claims versus H1: u<175 dollars in claims and a p – value equal to .3045. Using a significance level of .05, data shows p – value greater than .05. From this information I failed to reject Ho. I conclude that there is insufficient evidence that u < 175 dollars claimed. We can determine the effect size of this data by using Cohen's D calculation, which is: D equals the absolute value of -6.58554 dollars claimed divided by 203.37989 dollars claimed, which equals, .0324 dollars claimed. This shows the effect size being very small. This is not practically significant.

The study conducted by the appliance company shows that the company was trying to determine whether charging customers 175.00 dollars for a warranty on a specific brand of refrigerator made the company a profit. We found that the data proved that the company did not spend on average less than 175.00 dollars on a claim. The effect of the warranty in regards to profitability for the company is very small. We can infer that the warranty price needs to be higher or that the data collection procedures need to be more detailed to recognize where profitability, if any, lies.

The appliance company should have hired a professional company to conduct this survey. This is way too important for the company's finances to not invest in having the most accurate data with correct collection procedures. I feel that the brand of refrigerators should have been broken down into the specific models so the company could see which models were having more claims and which models were more dependable. This would allow the company to either discontinue selling the model in the store or not offer a warranty on that particular one.

In conclusion, we see that the research question, "Does an appliance company make a profit when they offer a two year warranty on a certain brand of refrigerator" does correlate with the data but because of flaws in the data collection procedure the data is unreliable.

TASK 4

The Research question, "Is there racial bias in sentencing for felonies", does correlate with the study that has been conducted, but there are some data collection procedures that could have flawed the results.

The study was conducted by an unknown Criminologist. We don't know his or her race or who he/ she is affiliated with. If this criminologist wants his research to be credible and unbiased then he should use a reputable, professional company to conduct his study. This would give his study the backbone it needs to establish results that are non-bias and accurate. This ensures credibility for other researchers who might want to use the study's data or simply to educate the public. The study is an observational study where the data was acquired from a database of criminal offenses in one state. Ideally, the study should have used several different randomly picked states that used records of randomly picked offenders of different felonies in the past year. This would have better represented the population of all sentenced felons.

The units of measurement in the study were used appropriately by using the length of sentencing in months. The mean for group 1 (Caucasian race) is 75.5492 months and the mean for group 2 (non-Caucasian) is 77.7680 months. There is not a large center difference between the 2 groups. The standard deviation for group 1 is 8.22424 months and for group 2, 7.10535 months. This data shows group 1 having a greater distance from the mean than group 2. The distribution for Caucasians (group 1), shown by the histogram, is right skewed and we see a similar pattern in Non-Caucasians (group 2) with a right skewed histogram. We see a small difference in the box plots of the two groups. Group 1 (Caucasians) showing a minimum data point of 68 months and a maximum data point of 116 months, while group 2(Non-Caucasians) showing a minimum data point of 71 months and maximum data point of 100 months. The range of group 1 is larger, with 48 months compared to the range of group 2 with 29 months. This difference includes 4 outliers for Caucasians (group 1) and 1 outlier for Non-Caucasians (group 2).

The T- test can be done because both groups have a large sample size equal to 30 or larger. When conducting a hypotheses test for this study we use, H_0: $\mu_1 = \mu_2$ vs. H_1: $\mu_1 < \mu_{2.}$ The P-value for the test is “Sig. (2-tailed) value” on the SPSS output divided by 2, which calculates, .024 divided by 2, which equals .012. Using a significance level of .05 and if p-value is < .05 then reject $H_{o.}$ We reject H_o. There is sufficient evidence to conclude that Caucasian offenders will get more lenient sentences than any other Non-Caucasian groups. This result is statistically significant. The effect size of the two-sample test is the Cohen D test. It is calculated by using the formula: $d = \frac{|\bar{x}_1 - \bar{x}_2|}{s}$. We use group 1 sample mean of 75.55 months minus 77.77 months for group 2 and the absolute value is 2.22 months divided by the larger standard deviation of the 2 groups, which is 8.22424. Our effect size is .2699, which is small and is not practically significant.

In this study we see that just because a hypotheses is statistically significant doesn’t mean that your results are practically significant. In this situation, I feel the large sample amount of one type of felony and the sample collection of offenders from just one state, contributes to the small effect size. Even though Non-Caucasians have a higher average mean for sentences received, those sentences were not high enough over Caucasians mean to practically support the study. Sentencing can be so varied, for example; first time offenders may not get as much time as second time offenders, secondly, different felonies carry different sentencing structure and time, states have different sentencing structure and time, and lastly, sentencing can depend on how violent the crime is (weapons involved, injuries sustained by the victim, etc.). The point is, that taking a sample of one particular type of felony (burglary) from one state does not represent the population of this study very well, which affects the results. The study also shows four outliers for Caucasians, this could have affected the results. The Criminologist may want to include the outliers and exclude the outliers in the study and compare data to see, if any, the affect they may have on the results. The one outlier for Non-Caucasians can probably be assumed an error.

As we can see, the research question, “Is there racial bias in sentencing for felonies?” does correspond with the data presented in this study. Unfortunately, the data collection procedures were flawed which made the results unreasonable. The study determined that there was sufficient evidence to conclude that Caucasian offenders get more lenient sentences than any other group (non-Caucasians). These results were statistically significant but when we conducted the Cohen’s D test, the effect size was small which made this study not practically significant. I feel if the study had used samples from randomly selected states and randomly selected felonies, other than just burglaries, it would have given more believable data results with possibly a larger effect size, which would make this study practically significant as well. If the Criminologist had used a reputable professional company to conduct his study, he may have had a more reasonable outcome.

APPENDIX E: FULL TEXT OF KURT’S STATISTICAL ARGUMENTS

TASK 1

A research study was done in order to find out whether or not teens are paying attention to local and national politics. Researchers recruited 50 teenagers at a popular arcade in a local mall and asked them the following question: “Would you support suffrage for 16- and 17-year olds?”. First of all it has to be said that the sampling and data collecting procedure does not completely fit the research question. There are several reasons which result in the assumption that the procedure is misleading and does not bring reliable results. On the one hand there is a possibility that teenagers want the right for suffrage just to have the feeling of being adults or they could also just seek attention and a feeling of power. So if they answer yes to the questions it does not necessarily mean that they pay attention to politics. On the other hand it could be a leading question to a certain degree. Teenagers probably tend to answer yes because they want to come across as “good” citizen. The interviewer might come across as some kind of authority which increases the teenagers’ motivation to lie. Another thing about the procedure which can be criticized is the limited variety of the sample group. The researchers just chose individuals from the same spot in a local mall. This is called convenience sampling and does lead to a biased sample. In order to get a representative sample they should have asked teenagers all across the nation and also not only in malls. It is important to get representative from all social ranks and from different states as well. It can be generally stated that the sample size is also not big enough. The researchers try to find out the opinion from a very large population. 50 teenagers are not enough to represent every teenagers in America, especially if the sample is biased to a certain degree.

The result of the research study is theoretically reasonable. 24 percent of the sampled individuals said that they support suffrage for 16- and 17-year olds and 76 percent are against it. It is generally known that teenagers are not really interested in politics. According to the criticism concerning the sampling and data collecting procedure it has to be assumed that the results of the study are not reliable.

Taking everything into consideration it can be stated that the sampling and data collecting procedure is not sufficient enough to find satisfying results. The question used for research is leading and does also not fit the research question completely. Furthermore the sample is not big enough to represent the population and in addition to that the sample is biased because of a convenience sampling method which leads to a lack in variety. Finally, according to the study the research question can be answered as follows: The majority of teenagers do not pay attention to local and national politics, but more data is required from a survey with a more sufficient sampling and data collecting procedure and also a sample which is representing the population before it can be determined.

TASK 2

A research study was done in order to find out whether or not drivers of red and black colored cars do generally drive with higher average speed than drivers of neutral colored cars. In order to find this out police officers were told to record the speed and color of passing cars in 60mph zones in different parts of the city at different times of the day. The data collection procedure does fit well to the research question. Furthermore the usage of the unit miles per hour makes perfect sense for the considered scenario. Police officers recorded 75 data points of different colored cars in total. The recorded the speed of 30 neutral colored cars, 23 black colored cars and 22 red colored cars. It can be generally said the sample size of either color is too small to receive reliable results. The researchers try to prove a general statement. In order to do so a very large sample size is required. In addition to that the sample sizes of the different colored cars should be equal. The research method in this scenario can be taken as a convenience sample. Police officers take speed controls at different parts of the cities and also at different times of the day, but since it's a general statement this is not enough variety. In order to achieve a high quality sample researchers should take data, collected from different cities and also in different speed zones, into consideration. Drivers are probably more likely to speed in different speed zones in comparison to others. To further investigate the validity of the research results and methods it is important to compare the center, spread and distribution of the findings.

In order to compare the center of the results for each car-type, the mean can be used as an appropriate measure. The graphs of the car-types are not tremendously skewed which justifies the usage of the mean in this scenario. Neutral colored cars have a mean of 65.96, black cars have a mean of 69.43 and red cars have a mean of 76.40. All measures represent miles per hour. According to the means neutral colored and black cars have lower average speed than red colored cars. Neutral colored car drivers drive the slowest average speed. Drivers of red colored cars drive around 16 mph over the speed limit which is remarkable. In order to find out the spread of the data of each car-type, the standard deviation has been taken into consideration. The standard deviation of black cars (3.43) and red cars (3.14) is similar. The degree of standard deviation is relatively low and proves that there is no remarkable variation in the collected average speed records of red and black cars. The standard deviation for neutral colored cars (5.98) is almost twice as high in comparison. This makes perfect sense in the given context. Neutral colored cars included a number of different colors which make the possibility for a spread higher. This underlines the statement that a higher sample size is needed for reliable results.

Last but not least there is the distribution of the data. The histogram of neutral colored cars is approximately bell-shaped, the histogram of black colored cars is left-skewed and the histogram of red colored cars is approximately bell-shaped again. The box plot of neutral colored cars has the biggest range and the lowest median. The box plots of the red and black colored cars are very similar in their range, just the median of the red colored cars is higher than the median of black colored cars. The data of neutral colored cars has one upper outlier at around 85 mph and the data of black colored cars has one lower outlier at around 61 mph. The fact that the graphs of the data of red and black colored cars is left-skewed means that there are more higher data points. So red and black cars are driven faster in average. The fact that the distribution of neutral colored cars is so high means that the data collected for those cars is not stating a specific trend. Some drivers of neutral cars are faster and others slower. The fact that there is an outlier at the data of black colored cars which is below the normal range underlines the fact that black colored cars are driven slower than red colored cars in average.

Taking everything into consideration the research question can be answered as follows: According to the collected data it can be stated that drivers of red and black cars are driving faster than drivers of neutral colored cars in average. Furthermore especially drivers of red cars are driving way above average speed, but the data collecting procedure must be adjusted to achieve satisfying and reliable results. The data collecting procedure is well chosen but there are some deficits. Things that can be improved in order to achieve a more reliable result are a bigger sample size, more variety in the collecting of data and collecting of data of another control group. Another control group could be for example green or yellow cars. Variety can be achieved by collecting data in different cities, countries and speed zones. In addition to that it should be exactly explained what colors are included in neutral colors. The study should have very good results if all these adjustments would be added to the procedure.

TASK 3

Research was done by an appliance company in order to find out whether or not the company makes a profit with a certain two-year warranty which they offer for refrigerators. Customers pay 175 dollars for this particular warranty, so the company only makes profit if the average cost of processing the customers' claims is less than 175 dollars. The sampling and data collecting procedure used for this research question is that they monitor the claims of the first 250 customers who purchase the warranty and see how high if the claims within two years are in total. The research question and data collecting procedure make sense, because the result should clearly show whether or not there is a profit. Another good feature of the data collecting procedure is that the sample they chose is a random sample. The first 250 customers could be from any race, age, gender and so on. Also the sample size is large enough to lead to reliable results.

In order to determine the center of the data it is appropriate to use the median instead of the mean. The histogram according to the data is strongly skewed which indicates a big difference between median and the mean and it also indicates that the median should be used. The median is about 118.05 dollars. The median looks promising for the company because the amount of claims is less than the revenue the company makes. Furthermore it is also important to look at the spread of the data. In order to do so the standard deviation has to be taken into consideration. The standard deviation is 203.3799 dollars. The standard deviation is fairly large, which is a negative factor for the reliability of the data collection. It basically means that there are probably a lot of outliers and that there are a lot of customers who have claims which are higher than the price they initially paid. The next factor which is important for the reliability of the data is the distribution. The distribution can be determined by having a look at the histogram, range, skewness and box plots. The box plot has seven outliers and three of them are extreme outliers. The extreme outliers range from about 800 to 880 dollars and the other outliers range from about 500 dollars to a little bit over 600 dollars. The histogram is strongly right skewed which indicates that low data points are most common. This explains why the median is fairly low even though there are a lot of higher outliers and a fairly high standard deviation. The regular range of the box plot pends from close over zero dollars to about 450 dollars. In terms of H0, μ is equal to 175 dollars and in terms of H1, μ is less than 175 dollars. The p-value obtained is 0.3045, which means that we cannot reject H0. There is no sufficient evidence that the average amount of claims that has to be covered by the company is less than 175 dollars. So there would not be a profit for the company. The effect size, calculated through Cohen's d, came out to be -0.0324 which is considered to be a small effect size.

Taking everything into consideration it can be initially said that the answer for the research question cannot be answered with yes. Most of the investigation towards the data analysis shows that the amount of claims tends to be higher than the revenue made by the company by selling the warranty in the first place. The only fact that speaks in favor of the company's pricing is the center of the data collection. The spread and also the distribution show the data is strongly varying, especially towards a higher amount. Outliers are only on the higher side of the box plot too. Finally, the hypothesis testing also measures that the average amount of claims is not below 175 dollars. There is no recommendation in order to improve the data collection procedure. The research was done fairly well. The company should overthink its pricing strategy again and maybe raise the price for a warranty or maybe shorten the period of time in which customers can refer to this warranty.

TASK 4

Researchers try to find out whether or not there is any kind of racial bias in sentencing for crimes. In order to find information to answer the question, criminologists used the data of a given state and compared the race of offenders with the length of their sentence. Burglaries are used at the crime in this sample. Due to previous studies, Caucasian people are found to be less likely to be sentenced biased. According to that study, our researches divided the sample members in two groups, Caucasian and Non-caucasian people. First of all, the research question does fit to the data collecting procedure, but there are also some deficits. The researchers in the particular case only focus on the act of burglaries. The degree of biased judgment could differ with the nature of the crime. In order to find results which cover crime sentences in general in connection with race, researchers should take a variety of different crimes into consideration. Another thing which has a negative influence on the reliability of the study is the fact, that the researchers only gathered data from one state. In order to answer the research question better, they should spread their investigation among more states or even countries. The sample size is big enough to find reliable results. They found over a hundred cases for both sample groups. So the sample size of both groups is big enough to notice a difference in judgment based on potential biases. The length of the crime sentence is measured in month, which makes perfect sense for this study. Another thing that can be criticized is that the category of Non-Caucasian people is really broad. There could be bigger biases against specific races, so it might be a good idea to proof races individually

In order to determine the center of the data it is appropriate to use the median instead of the mean. The histogram according to the data is strongly skewed which indicates a big difference between median and the mean and it also indicates that the median should be used. The median for group 1 (Caucasian people) is 73 month of sentence received, the median for group 2 (Non-Caucasian people) is 75 months of sentence received. According to the center of the data there is slight difference among the two groups. The median of group 2 is slightly higher, which does support the assumption that Non-Caucasian people are judged differently.

Furthermore it is also important to look at the spread of the data. In order to do so the standard deviation has to be taken into consideration. The standard deviation of group 1 is 8.2242 months of sentence received and the standard deviation for group 2 is 7.10535. The spread of the data is similar among both groups, but it can also considered to be high. 8 months of sentence are a long time, so in terms of the study it means that the data varies.

The next factors which is important for the reliability of the data is the distribution. The distribution can be determined by having a look at the histogram, range, skewness and box plots. The range of the data of group 1 (68-90) is much smaller than the range of group 2 (72-96). Group 1 has three outliers and one extreme outlier. All outliers are on the upper side of the box plot. The normal outliers range from 95 to 105 months of sentence received and the extreme outlier lies on approximately 117 moths of sentence received. Group 2 shows online one outlier on the upper side at around 100 months of sentence received. The histograms of both groups are strongly right-skewed, which indicates that low values are most common. Summarizing the distribution of the data of both groups it can be said that the length of the sentence varies a lot in both groups. The general length of sentences seems to be higher for group 2, but on the other hand group 1 has a lot of outliers. So according to the distribution group 2 has the higher length of sentences but in group 1 there are some individual cases with very high data points. This supports the claim of the researchers.

Due to the fact that the sample sizes are above 30 the assumptions have been met for both groups to conduct the t-test. The t-test helps us to investigate whether or not there is a difference in the judgment of different races because of potential biases. In terms of H0, μ1 is equal to μ2 and in terms of H1, μ1 is smaller than μ2. The p-value obtained is 0.012, which means that we can reject H0. There is sufficient evidence to conclude that Non-Caucasian people do get longer sentences for their crimes of burglary. The effect size, calculated through Cohen’s d, came out to be 0.2698, which is on the smaller end. It means that even though the hypothesis test came out to identify a difference between both groups, this difference can be considered to be small.

Finally it can be concluded that there is a difference between the two groups according to the findings. The center, spread and distribution do not really support the fact that there is a reasonable difference among the judgment of both groups. Only the hypothesis test supports the initial thesis and gives us reason to further investigate the connection between different races and their penalties for crimes. As already stated in the beginning, it is important to gather more data from different locations and to have more variety in terms of the nature of the crime.